Computational modeling of cohesin ATPase dynamics

Iñigo Marcos Alcalde

Directores de tesis:

Paulino Gómez Puertas y
Jesús Mendieta Gómez

Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM)

Programa de Doctorado en Física de la Materia Condensada, Nanociencia y Biofísica

Facultad de Ciencias

Universidad Autónoma de Madrid

Madrid, Febrero 2018

Agradecimientos

Agradezco a mis directores haberme ofrecido la oportunidad de realizar esta tesis y haberme guiado y enseñado pacientemente a lo largo de estos años. También agradezco a Jesús Ignacio y Fernando los buenos ratos que hemos pasado y las innumerables veces que me han ayudado.

Esta tesis está dedicada a mis padres, Ajo y el resto de mi extensa y maravillosa familia, por su cariño y todo lo que me han enseñado y a Lucía por acompañarme siempre y alegrarme cada día.

Resumen

El complejo de cohesina humano participa en procesos de segregación cromosómica, reparación de ADN, organización de cromatina y regulación de la transcripción. Este complejo, que forma un anillo que atrapa topológicamente DNA, está formado por cuatro subunidades principales: Smc1A, Smc3, Rad21 y Stag1/2. Determinadas variantes de componentes de este complejo que se han relacionado con cohesinopatías y cáncer. En esta tesis se han empleado una serie de métodos de biología estructural computacional (modelado por homología, técnicas de dinámica molecular clásica, dirigida y de potenciales híbridos de mecánica cuántica y mecánica clásica, así como docking molecular) para elucidar aspectos relevantes de la dinámica de cohesinas vinculada a la actividad ATPasa. Además, se ha desarrollado MEPSA, una herramienta accesible que permite estandarizar el análisis de superficies de energía libre tridimensionales (un tipo de dato que resulta frecuente en múltiples protocolos de dinámica molecular). Los resultados obtenidos en esta tesis han permitido: i) describir cómo la unión del dominio C-terminal de Rad21 a la cabeza de Smc1A facilita la actividad ATPasa en el centro activo de Smc1A en el anillo de cohesina, ii) caracterizar que esta actividad ATPasa desencadena un proceso de acoplamiento alostérico entre los centros activos de Smc1A y Smc3 no descrito previamente y iii) evaluar el efecto desestabilizador que tiene la hidrólisis de ATP en ambos centros activos sobre la interfaz formada por los dominios cabeza de Scm1A y Smc3. Asimismo, la descripción a escala atómica que ofrecen los modelos generados ha permitido racionalizar múltiples variantes relacionadas con enfermedades humanas (síndrome de Cornelia de Lange y cáncer) y mutaciones no neutras en levadura, así como la detección de una nueva diana molecular para la búsqueda computacional de fármacos. Los resultados preliminares de un protocolo de docking molecular contra dicha diana han permitido identificar una molécula capaz de inducir arresto en fase G2/M del ciclo celular en la línea celular humana 293T.

Summary

Human cohesin complex is involved in processes of chromosome segregation, DNA repair, chromatin organization and transcription regulation. This complex, which forms a ring capable of topologically entrapping DNA, is formed by four core subunits: Smc1A, Smc3, Rad21 and Stag1/2. Variants affecting members of this complex have been directly related to cohesinopathies and cancer. In this thesis a series of computational structural biology methods (homology modeling, classical, biased and hybrid quantum mechanics/molecular mechanics molecular dynamics techniques, as well as molecular docking) have been used to elucidate relevant aspects of cohesin ATPase dynamics. In addition, a user-friendly free energy analysis software (MEPSA) has been developed, streamlining the analysis of 3D free energy surfaces (a common type of data generated in many molecular dynamics protocols). The results presented in this thesis have allowed to: i) describe how Rad21 C-terminal domain binding to the Smc1A head domain facilitates ATPase activity in the Smc1A active site of a cohesin ring, ii) characterize that this ATPase activity triggers a previously unreported allosteric coupling mechanism between Smc1A and Smc3 active sites and iii) quantify the destabilizing effect ATP hydrolysis in both active sites has over the interface formed by Smc1A and Smc3 head domains. This structural framework has made it possible to rationalize many disease-related (Cornelia de Lange syndrome and cancer) human variants and yeast non-neutral mutations, also revealing a new molecular target for *in silico* drug discovery. Preliminary results of a molecular docking protocol against this target have yielded a drug capable of inducing G2/M cell cycle arrest in the 293T human cell line.

Contents

List of Articles

A list of the articles published during the development of this thesis is presented bellow.

Currently submitted:

- S. Gudmundsson, G. Annéren, I. Marcos-Alcalde, M. Wilbe, M. Melin, P. Gómez-Puertas & M. L. Bondeson. Novel RAD21 p.Gln592del variant expands the clinical description of Cornelia de Lange syndrome type 4 – review of the literature. (Submitted)
- B. Puisac, I. Marcos-Alcalde, M. Hernández-Marcos, A. Levtova, B. Schwahn, C. DeLaet, B. Lace, P. Gómez-Puertas & J. Pié. Human mitochondrial HMG-CoA synthase deficiency: role of enzyme dimerization surface and characterization of three new patients. (Submitted)

2017:

- I. Marcos-Alcalde, J. I. Mendieta-Moreno, B. Puisac, M. C. Gil-Rodríguez, M. Hernández-Marcos, D. Soler-Polo, F. J. Ramos, J. Ortega, J. Pié, J. Mendieta & P. Gómez-Puertas. Two-step ATP-driven opening of cohesin head. *Sci Rep*, 7(1):3266, June 2017.
- R. M. Buey, D. Fernández-Justel, I. Marcos-Alcalde, G. Winter, P. Gómez-Puertas, J.M. de Pereda & J.L. Revuelta. A nucleotide-controlled conformational switch modulates the activity of eukaryotic IMP dehydrogenases. *Sci Rep* 7(1), 2648, June 2017.

2015:

- I. Marcos-Alcalde, J. Setoain, J. I. Mendieta-Moreno, J. Mendieta & P. Gómez-Puertas. MEPSA: minimum energy pathway analysis for energy landscapes. *Bioinformatics*, 31(23):3853-5, December 2015.
- J. I. Mendieta-Moreno, I. Marcos-Alcalde, D. G. Trabada, P. Gómez-Puertas, J. Ortega & J. Mendieta. A Practical Quantum Mechanics Molecular Mechanics Method for the Dynamical Study of Reactions in Biomolecules. *Adv Protein Chem Struct Biol*, 100:67-88, 2015.
- M. C. Gil-Rodríguez, M. A. Deardorff, M. Ansari, C. A. Tan, I. Parenti, C. Baquero-Montoya, L. B. Ousager, B. Puisac, M. Hernandez-Marcos, M. E. Teresa-Rodrigo, I. Marcos-Alcalde, J. J. Wesselink, S. Lusa-Bernal, E. K. Bijlsma, D. Braunholz, I. Bueno-Martinez, D. Clark, N. S. Cooper, C. J. Curry, R. Fisher, A. Fryer, J. Ganesh, C. Gervasini, G. Gillessen-Kaesbach, Y. Guo, H. Hakonarson, R. J. Hopkin,

M. Kaur, B. J. Keating, M. Kibaek, E. Kinning, T. Kleefstra, A. D. Kline, E. Kuchinskaya, L. Larizza, Y. R. Li, X. Liu, M. Mariani, J. D. Picker, A. Pie, J. Pozojevic, E. Queralt, J. Richer, E. Roeder, A. Sinha, R. H. Scott, J. So, K. A. Wusik, L. Wilson, J. Zhang, P. Gomez-Puertas, C. H. Casale, L. Strom, A. Selicorni, F. J. Ramos, L. G. Jackson, I. D. Krantz, S. Das, R. C. Hennekam, F. J. Kaiser, D. R. FitzPatrick & J. Pié. De novo heterozygous mutations in SMC3 cause a range of Cornelia de Lange syndrome-overlapping phenotypes. *Hum Mutat*, 36(4):454-62, April 2015.

2013:

- F. Martín-Garcia, J. I. Mendieta-Moreno, I. Marcos-Alcalde, P. Gómez-Puertas, and J. Mendieta. Simulation of catalytic water activation in mitochondrial F1-ATPase using a hybrid quantum mechanics/molecular mechanics approach: an alternative role for beta-Glu 188. *Biochemistry*, 52(5):959-66, February 2013.

List of Abbreviations

- ΔG°: Free energy difference.
- $\Delta^{\ddagger} G^{\circ}$: Free energy of activation.
- AS1: Active Site 1.
- AS1-ADP/AS2-ATP: simulation conditions of the cohesin head complex in which active site 1 was binding ADP and active site 2 was binding ATP.
- AS1-ATP/AS2-ATP: simulation conditions of the cohesin head complex in which both active sites were binding ATP.
- AS2: Active Site 2.
- ABC: ATP Binding Casette.
- BLAST: Basic Local Alignment Search Tool.
- CdLS: Cornelia de Lange Syndrome.
- DFT: Density Functional Theory.
- GB: Generalized Born.
- GUI: Graphical User Interface.
- HMM: Hidden Markov Model.
- MD: Molecular Dynamics.
- MEPSA: Minimum Energy Path Surface Analysis.
- MM: Molecular Mechanics.
- MSA: Multiple Sequence Alignment.
- NAOs: Numerical Atomic-like Orbitals.
- NCBI: National Center for Biotechnology Information.
- NMR: Nuclear Magnetic Resonance.
- P: Product.
- PBC: Periodic Boundary Conditions.
- PBE: Poisson Boltzmann Equation.
- PDB: Protein Data Bank.

- Pfam: online database of protein domain families.
- P-loop: Phosphate binding loop.
- PMEMD: Particle Mesh Ewald Molecular Dynamics.
- Rad21-Cter: C-terminal domain of human Rad21.
- RC: Reaction Coordinate.
- RC1: Reaction Coordinate 1.
- RC2: Reaction Coordinate 2.
- RMSD: Root Mean Square Deviation.
- RMSF: Root Mean Square Fluctuation.
- S: Substrate.
- Smc1A-head: human Smc1A ATPase head domain.
- Smc3-head: human Smc3 ATPase head domain.
- SMD: Steered Molecular Dynamics.
- SMC: Structural Maintenance of Chromosomes (protein family).
- SMTL: SWISS-MODEL Template Library.
- TS: Transition State.
- TST: Transition State Theory.
- QM: Quantum Mechanics.
- QM/MM: Quantum Mechanics/Molecular Mechanics.

List of Figures

List of Tables

1 Introduction

1.1 Overview

Living organisms show robust mechanisms that allow them to perform their functions, keeping strong levels of organization, both in space and time, in a wide range of conditions. These mechanisms rely on a series of well-orchestrated chemical reactions that have to take place in localized regions, at controlled rates and in a concerted fashion[1]. These three aspects can be successfully governed thanks to the action of enzymes (i.e. biological catalysts), the activity of which can be regulated through a wide range of mechanisms (e.g. compartmentalization, pH, cofactors, regulatory proteins, and other regulatory molecules). In order to understand and possibly alter or engineer biological processes (e.g. novel therapeutic approaches, industrial uses, etc.) we unavoidably need to describe in detail the catalytic mechanisms of enzymes, their regulatory mechanisms, the effect of their activities on their conformational dynamics, as well as the outcome of their interaction with other molecules. All these aspects are determined by the constituent atoms of each enzyme and the interactions between them, the substrates and the environment (e.g. the solvent, ions, other cellular components, etc.). Therefore, if mechanistic understanding of these systems is to be achieved, it must eventually be formulated in atomistic terms[2].

The most common notion of enzymes may depict them as catalysts which can accelerate certain reactions in order to increase or decrease the concentration of particular molecules in a given region and/or period of time. However, many proteins exhibit enzymatic activity, i.e. are enzymes, but their ultimate goal is not merely to transform their substrates into products but to couple these chemical reactions with conformational changes that ultimately exert their molecular function. Ubiquitous examples of this behavior can be found along the ABC (ATP Binding Casette) ATPase protein family (Pfam: PF09818)[3], which hydrolyze ATP to energize various biological processes.

In this thesis we will focus on the cohesin protein complex, in particular on its ATPase head region, which is structurally similar to an ABC ATPase transporter[4,5]. Cohesin is a protein complex that forms a ring which, among other functions, holds together sister chromatids during cell division ensuring a balanced chromosome segregation[6]. Opening of the ring, necessary for both loading and unloading of DNA, is regulated by the ATPase activity of a heterodimeric region, but there are still many relevant questions regarding the mechanistic understanding of this process that are yet to be addressed[7]. In addition, mutations in members of this complex are related with rare developmental diseases[8–13], the most frequent of which is Cornelia de Lange Syndrome (CdLS), as well as several types of cancer[14–22]. Unfortunately, in most of them, the molecular mechanisms that lead

into disease still remain unexplained.

Although molecular knowledge of the cohesin function is steadily increasing, atomistic description leading to full mechanistic understanding is yet to be achieved. To reach this level of characterization, information coming from molecular biology experiments has to be complemented by methods exhibiting atomic resolutions.

The dramatic increase in computation performance over the last decades, together with the development of novel algorithms, is positioning computational modeling as a fundamental workhorse in the atomistic description of biological processes. By making use of the available structural information through these computational techniques, the aims of this thesis were i) to obtain an atomistic description of the ATPase region of this complex to help answering some of those open questions, ii) offer a new scheme to rationalize disease causing mutations and iii) pave the way for the proposal of potential novel therapies. During the completion of these objectives, a free energy surface analysis tool, MEPSA (Minimum Energy Path Surface Analysis), was developed to facilitate the data analysis of the quantum mechanical simulations used to study the ATPase activity of cohesin. Altogether, the results of these efforts got condensed into two peer-reviewed publications[23,24] available in appendices A and B, around which the results section is structured, giving an extended vision on their hypotheses, methodologies and results. Prior to the description of those results, an introduction to the computational techniques used and a schematic depiction these thesis objectives are offered in sections 1.2 and chapter 2 respectively.

1.2 Computational Techniques in Structural Biology

1.2.1 Multiple Sequence Alignment

Variability of phenotypically relevant biological sequences (DNA, RNA or protein) is constrained by natural selection. Those sequences coding for key elements of organism fitness tend to vary significantly less and, as long as they keep being functionally relevant, they are usually conserved after both intra-genomic duplication and speciation phenomena. By exploiting such function dependent conservation, the search for homology relationships (i.e. sharing a common sequence ancestor) among biological sequences proves a powerful bioinformatics tool to tackle many typical problems in molecular biology[25].

Homology search requires comparison among sequences, which is performed by aligning those sequences while following scoring criteria weighting the conserved matches, point mutations, insertions and deletions. Given a pair of sequences and a scoring function, Needleman-Wunsch and Smith-Waterman dynamic programming algorithms provide the optimal global and local alignments respectively[25]. However, despite being inherently generalizable to multiple sequence alignment (MSA), the computational cost of classic dynamic programming algorithms grows exponentially with the number of sequences (N) in at least: $O(2^N L^N)$, where L is the average sequence length. This made such algorithms unpractical for almost any MSA thus forcing the development of faster heuristics capable of handling large number of sequences that approximate, but no longer guarantee, an optimal solution, being the progressive alignment algorithms the first to appear and most widely used[26].

The main characteristic of progressive alignment algorithms is the generation of a guide tree that is iteratively followed to build the final MSA in order to reduce computational complexity. The guide tree is originally generated by hierarchical clustering of a distance matrix of all the possible pairwise alignments. As these two steps (i.e. distance matrix calculation and clustering) can represent computational bottlenecks when dealing with a large number of sequences, different heuristic solutions have been developed[26]. Specifically, during this thesis Clustal Omega, the latest version of the widely used Clustal MSA package, was used[27–29]. The progressive alignment algorithm used in Clustal Omega is capable of performing fast and accurate MSAs of a large number of sequences. By default, instead of calculating all pairwise alignments among all N input sequences, Clustal Omega uses a modified version of the mBed algorithm. If less than 100 input sequences are given, mBed calculates a full distance matrix but, if more than 100 input sequences are used, mBed only calculates the pairwise distances of all N sequences against a $(\log_2(N))^2$ long subset of randomly chosen seed sequences. Alignments are performed with fast k-tuple methods and the resulting distances are stored in N vectors, one for each input sequence, of n elements, one for each seed sequence. These vectors are quickly clustered by an iterative bisecting k-means algorithm, having a hard-coded 100 elements cluster size limit. To build the final MSA, sequential profile-profile alignments along the resulting guide tree must be performed, generating larger and larger MSAs until all the sequences are contained. In Clustal Omega, profiles are converted by default to Hidden Markov Models (HMMs) and aligned through the HHalign package.[27,30–32] The rough description of the Clustal Omega workflow depicted in this introduction refers only to the default and more common options, as it has been used in this work. However, for particular tasks, many

aspects of the algorithm behavior can be tuned, activated or deactivated through a wide range of command-line options and/or different input files.

1.2.2 Homology Modeling

Homology modeling is the most widely used and accurate method for 3D structure prediction. It is based on the observation that, in general, evolutionarily related protein domains tend to share similar 3D shapes. Interestingly, 3D homology relationships tend to last longer than sequence and function similarity and, therefore, it is considered the most robust conserved feature of homologous protein domains. Although there are some exceptions to this notion[33] it is still valid for most of the cases[34]. Therefore, generally having a protein with known 3D structure should provide information about the 3D shape of its homologous proteins detected by the more labile sequence-sequence similarity, at least for the highly conserved regions.

Thus, to obtain a structural model for a protein sequence, if we can detect homology relationship by sequence similarity to another protein sequence with known 3D structure, we should be able to assume a similar 3D shape for, at least, the sequence conserved regions, obtaining a 3D homology model. The closer the homology between sequences is the more reliable a model can be expected to be. This implies that good quality alignments, capable of properly capture the conservation of key patterns and structures, are an essential requirement to obtain insightful homology models.

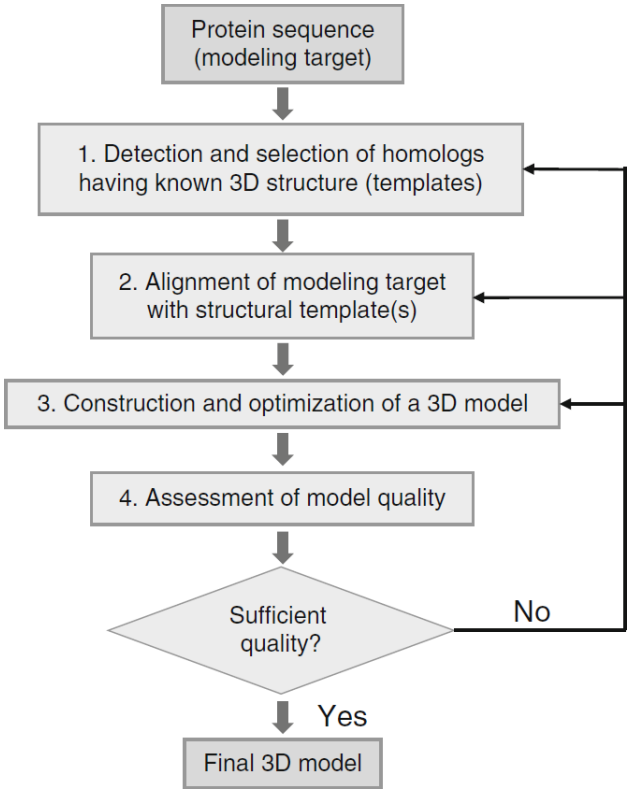


Figure 1: *Homology modeling flowchart (Source: Venclovas, 2012[34]).*

Any homology modeling workflow (Fig. 1) consists of four iterative main steps[34]:

- Identification of homologous sequences with experimentally solved 3D structure to be used as the template.
- Sequence-structure alignment to map the equivalent residues between target and template sequences.
- 3D model generation using the template 3D structure and the sequence-structure alignment.
- Model quality evaluation and workflow repetition until the quality threshold is reached.

In this thesis, these steps were taken as follows. Identification of homologous sequences with known 3D structure was performed using a Basic Local Alignment Search Tool (BLAST)[35] search through the US National Center for Biotechnology Information (NCBI) web servers[36], using the sequence to be modeled as query and Protein Data Bank (PDB) as the target database[37]. Once a homology candidate was detected, the conservation of each sequence region was refined by generating an MSA of closely homologous protein sequences as well as our target and template sequences with Clustal Omega[27–29] that was manually curated afterwards. Sequence-structure alignment was extracted from the resulting MSA as, when working with closely related sequences, the MSA resulting from this procedure usually offers an accurate sequence-structure alignment[34].

3D models were generated via SWISS-MODEL[38] using the sequence-structure alignments obtained with the described procedure and the template structure.

SWISS-MODEL is an automated homology modeling web server which, starting from an input protein sequence, can automatically select a template or templates from the SWISS-MODEL template library (SMTL), build a model from it/them and then evaluate its quality. SMLT consists on a curated and annotated library of imported PDB[37] entries and, storing their associated template sequences in BLAST[35] searchable databases as well as in HHblits[39] searchable HMM libraries[38]. When a target protein sequence is introduced, SMLT is searched both with BLAST and HHblits obtaining a series of target-template alignments which are then filtered by a wide range of criteria (sequence identity, sequence similarity, HHblits score, agreement between predicted secondary structure and solvent accessibility of target and template)[38]. If more than one template is found a structurally corrected MSA of their sequences is performed and an average framework is generated. Input sequence is added to this MSA generating the alignment, obtaining the final alignment that describes residue correspondence[40]. The automation described up to this point is optional as this process can be manually performed via the DeepView program[40], allowing manual curation of the alignment, which can be worthy for the more complex modeling tasks, as is the case of the models generated during this thesis. Either starting from the fully automated process or a DeepView project, once a sequence-template alignment is obtained, the SWISS-MODEL server derives atomic coordinates from the structural framework following the sequence-template alignment. Then nonconserved loops are modeled, missing backbone and side chains are reconstructed and an estimation of the model quality is provided by the QMEAN composite scoring function[38].

In our case QMEAN was used just as a first pass filter, as 3D models were stabilized through Molecular Dynamics simulations monitoring the root mean square deviation of their alpha carbon traces to converge at least to be in the range of the experimental

resolution of the template structures. When excessive instability was detected, models were discarded, iterating through the modeling procedure until stability was finally reached.

1.2.3 Atomistic simulations

Atomistic description of biomolecular phenomena is often experimentally unreachable and can only be addressed through computer based simulations. Simulations can be performed so that this atomistic description may lead to explanations of macroscopic events, yet still retaining the possibility of discerning distinct levels of relevance between the implicated atoms or molecules. This differential relevance can prove highly valuable in proposing molecular biology experiments (e.g. proposal of novel drugs or prediction of phenotype-inducing mutations) from an atomistic perspective, which is experimentally unattainable. Thus, these techniques offer a powerful tool to unravel previously uncharacterized molecular mechanisms, especially when the conclusions they offer can be experimentally validated afterwards. The wide range of time (from femtoseconds to seconds) and length scales (from angstroms to tens or hundreds of nanometers) in which biomolecular phenomena take place makes the requirements of atomistic biomolecular simulations highly dependent on the particular process to be studied. In this section, the atomistic simulations techniques used during this thesis will be introduced.

1.2.3.1 Molecular dynamics

In molecular dynamics (MD) simulations, given an atomistic system and a potential function, Newton's second equation of motion is numerically solved over discrete time steps for every single atom, obtaining a description of the dynamics of that system in terms of positions and momenta of its atoms over time. There is a plethora of well-established MD packages available, e.g. NAMD[41], GROMACS[42], AMBER[43], CHARMM[44], etc. In this thesis, due to previous expertise in the group, AMBER package[43] was used. AMBER refers to two things, a package of MD simulation programs and a molecular force field family (which can be used with other MD simulation packages). In this thesis both the MD package and the force field were used to generate the MD simulations and, therefore, both aspects will be commented in this introduction.

Sander was the first AMBER module capable of performing energy minimizations and MD simulations among other functionalities. The most frequently used methods of sander were rewritten with the objective of improving their performance in a reimplementation called PMEMD (Particle Mesh Ewald Molecular Dynamics). One of the latest features implemented in this regard is the ability to use Nvidia GPUs to parallelize calculations. This GPU compatible executable was used in all the MD simulations generated during this thesis, except when a hybrid quantum mechanics/molecular mechanics potential was used, in which case a modified version of sander was used.

Molecular mechanics force field

For a simple isolated atomistic system, the second law of motion can be written as:

$$m_i a_i = f_i \qquad\qquad f_i = -\frac{\delta U}{\delta r_i} \tag{1.1}$$

The forces f_i acting over each atom (i) with r_i coordinates can be derived from the given potential energy function U. Once the corresponding force is obtained, as the mass of each atom (m_i) is known, acceleration (a_i) can be calculated. This potential energy, over which forces are calculated, is function of the atomic positions and is usually defined by a molecular mechanics (MM) force field. Although the nature of atomic interactions is essentially quantum mechanical, the poor scalability of quantum mechanics methods make these prohibitive for the kind of simulations MD is designed to tackle (10^5 or more atoms during time scales of nanoseconds to milliseconds). Therefore, simpler approximations, still capable of describing atomic motions accurately enough, are used (e.g. MM force fields). Depending on the particular system and the properties to be studied, the information the potential function has to describe varies. A potential function has to offer a sufficiently accurate description of atomic interactions and, yet, be inexpensive enough to make meaningful time scales and system sizes computationally affordable. Therefore there is no universal force field but rather a wide range of individual developments with different aims. In this section we will focus on the AMBER force field[45,46], which is widely used for the simulation of DNA and proteins. It is accurate enough to describe the motion of atoms in biomolecules when there is no bond breaking or formation, which is perfectly suited to study conformational dynamics and protein-protein interactions. Later in the "Fireball" subsection of this chapter, the more accurate, yet far more expensive, quantum-mechanical potential used to study chemical reactions will be discussed.

To allow the simulation of large polymers as proteins or nucleic acids, AMBER force field exhibits a simple enough potential energy function which offers good balance between sampling capabilities and accuracy in this type of systems. This potential energy function can be decomposed in bonding and non-bonding terms:

$$U = U_{bonding} + U_{non-bonding} \tag{1.2}$$

In the bonding potential energy ($U_{bonding}$) term, covalent interactions, which has to be specified in the parameter-topology file (see "Input files" subsection), are defined via harmonic potentials constraining distances (bonds), bend angles and dihedral angles (proper and improper torsions):

$$U_{bonding} = U_{bonds} + U_{angles} + U_{dihedral} \tag{1.3}$$

The potential energy for the bonds (U_{bonds}) can be written as:

$$U_{bonds} = \sum_{i}^{bonds} k_i^r (r_i - r_{i,eq})^2 \tag{1.4}$$

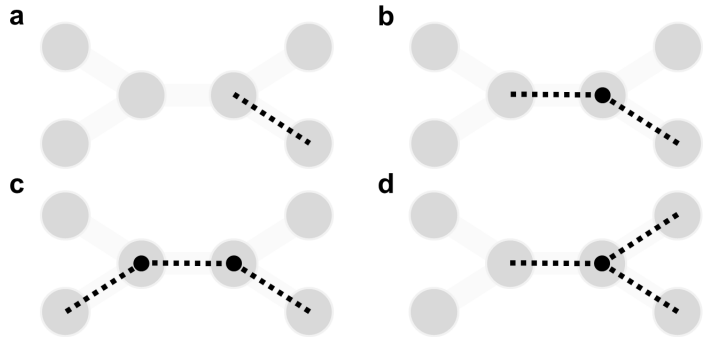


Figure 2: *Schematic representation of the bonding potentials defined in the force field: distances (**a**), angles (**b**), proper torsions (**c**) and improper torsions (**d**).*

where r_i is the distance between atoms forming the bond (Fig. 2 a), $r_{i,eq}$ is the equilibrium distance and k_i^r is the harmonic potential constant, both specified in the force field for each bond.

The potential energy for the angles (U_{anlges}) can be written as:

$$U_{anlges} = \sum_{i}^{angles} k_i^{\theta}(\theta_i - \theta_{i,eq})^2 \tag{1.5}$$

where θ_i is the angle value (Fig. 2 b), $\theta_{i,eq}$ is the equilibrium angle and k_i^{θ} is the harmonic potential constant, both specified in the force field for each angle.

The potential energy for dihedral angle definitions ($U_{dihedral}$) describing proper and improper torsions can be written as:

$$U_{dihedral} = \sum_{i}^{dihedrals} k_i^{\phi}[1 + \cos(\phi_i - \phi_{i,eq})] \tag{1.6}$$

where ϕ_i the value of the dihedral angle, $\phi_{(i,eq)}$ is the equilibrium angle and k_i^{ϕ} is the harmonic potential constant, both specified in the force field for each dihedral angle. When the dihedral angle is measured over consecutively bonded atoms, it is considered a proper torsion (Fig. 2 c), in contrast with improper torsions (Fig. 2 d), in which the dihedral angle formed by non-consecutively bonded atoms is measured.

Non-bonding energy ($U_{non-bonding}$) describes van der Waals and electrostatic contributions:

$$U_{non-bonding} = U_{LJ} + U_{Coulomb} \tag{1.7}$$

Usually van der Waals forces are modeled through Lennard-Jones (U_{LJ}) potential and electrostatic interactions via Coulomb's law ($U_{Coulomb}$) which, for ij pairs of atoms, can be written as:

$$U_{LJ} = \sum_{i<j}^{atoms} 4e_{i,j}\left[\left(\frac{\sigma_{i,j}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{i,j}}{r_{ij}}\right)^{6}\right] \tag{1.8}$$

$$U_{Coulomb} = \frac{1}{4\pi\epsilon_0}\sum_{i<j}^{atoms}\left[\frac{q_i q_j}{r_{ij}}\right] \tag{1.9}$$

where r stands for the distance between atom pairs, q for the charge of each atom and ϵ_0 for the vacuum permittivity constant. In Lennard-Jones potential definition e is the depth of the potential well that describes the interaction between atoms i and j and σ is the distance between atoms i and j at which the potential is zero, or, in other words, the distance at which the interaction between atoms i and j starts to be repulsive. Both e and σ have to be defined fore each atom pair in the force field. As both $U_{non-bonding}$ terms have to be calculated for every ij pair of atoms, the computational complexity for computing $U_{non-bonding}$ for N atoms is $O(N^2 + N)$. Therefore, even when fast evaluation methods are used, the impact this calculation has over performance is dominant over $U_{bonding}$ and makes the simulation of biomolecular systems, which usually consist of several thousand atoms, computationally intractable. To address this problem distance cut-offs are typically applied so that, for each atom, only the non-bonding interactions with atoms closer than the specified cut-off are calculated. In the case of the simulations performed during this thesis, the non-bonding cut-off used was 12 Å.

Summing all the previously mentioned terms, the potential energy function of the AMBER force field can be written as:

$$\begin{aligned} U_{bonds} = &\sum_{i}^{bonds} k_i^r(r_i - r_{i,eq})^2 + \sum_{i}^{angles} k_i^\theta(\theta_i - \theta_{i,eq})^2 \\ &+ \sum_{i}^{dihedrals} k_i^\phi[1 + \cos(\phi_i - \phi_{i,eq})] \\ &+ \sum_{i<j}^{atoms}\left[\left(\frac{A_i A_j}{r_{ij}}\right)^{12} - \left(\frac{B_i B_j}{r_{ij}}\right)^{6}\right] + \frac{1}{4\pi\epsilon_0}\sum_{i<j}^{atoms}\left[\frac{q_i q_j}{r_{ij}}\right] \end{aligned} \tag{1.10}$$

Thermodynamic ensemble

MD simulations are always performed in a particular thermodynamic ensemble, depending on the process to be studied or the conditions that are to be simulated. PMEMD implementation supports three different thermodynamic ensembles:

NVE: This ensemble, also known as microcanonical ensemble, represents a fully isolated system with fixed volume in which no heat or particles are exchanged with the environ-

ment. In this configuration the number of particles (N), volume (V), and total energy (E) remain constant.

NVT: This ensemble, also known as canonical ensemble, represents a system with fixed volume and no particle transfer with the environment which is immersed in a heat bath much larger than the system. The system is constantly in thermal equilibrium with the heat bath and, therefore, if the temperature of the system changes due to any internal process, a heat flux from or to the bath is established, restoring the temperature of the system to that of the bath. In this configuration the number of particles (N), volume (V) and temperature (T) remain constant.

NPT: This ensemble, also known as isothermal-isobaric ensemble, represents a system with fixed pressure and no particle transfer with the environment which is immersed in a heat bath much larger than the system. Similarly to NVT, the heat bath implies that any internal temperature fluctuation of the system will be absorbed by the bath. In this configuration the number of particles (N), pressure (P) and temperature (T) remains constant. This ensemble is the one that best mimics experimental conditions in an enzymology laboratory and, therefore, is the one used in the simulations generated during this thesis.

In MD, in order to control pressure and temperature, barostat and thermostat algorithms have to be used. In this thesis, temperature was regulated via a weak-coupling algorithm[47], which applies a scaling factor to the velocities of all atoms proportional to the difference between the system temperature and the reference temperature value (ntt=1 option in AMBER). Pressure was controlled with an isotropic Berendsen barostat[47] that periodically rescales all coordinates by a factor proportional to the difference between the system pressure and the reference pressure value (barostat=1 option in AMBER).

Initial structures

There are two typical sources of initial structures to perform MD simulations with biomolecules: structures that are available in the PDB[37], usually from X-ray crystallography or NMR (Nuclear Magnetic Resonance) experiments, and homology models. In either case, the election of a good initial structure from the PDB is crucial to obtain informative results from the MD simulation. Some of the factors to take into account during the election of an initial structure are the methodology with which the structure has been obtained, its resolution, how ligand locations have been determined, which conformation is stabilized with the ligand used, the backbone mobility (e.g. b-factor in crystallographic structures or conformer comparison in NMR) of certain regions, missing complex members, or completely unresolved regions. In the case of this thesis, homology models had to be used as there were no structures available in the PDB for the human ATPase head region of the cohesin complex bound to the C-terminal domain of Rad21. A description of the homology modeling procedure is available in section 1.2.2 and the details about the generation of the models used in this thesis can be found in section 3.2.4.

Protonation state and disulphide bond prediction

In MD simulations, protonation states of residues and disulphide bonds are fixed and they have to be determined during the preparation of the initial structure. There are some pKa calculation tools freely available like PROPKA[48], MCCE[49,50] and H++[51–53]. H++ was the one used to determine the protonation states of the structures used in the simulations performed during in this thesis. Disulphide bond presence was evaluated by manually checking the geometries of all the cysteine residues side chains present in the structure one by one. Thorough disulphide bond detection is particularly important as otherwise rearrangements of the tertiary structure of the protein could be artificially induced. If a disulphide bond is detected, it has to be defined via the LEaP module of the AMBER package to annotate it in the parameter-topology file (see "Input files" subsection).

Solvent model

Biomolecules are usually surrounded by aqueous environments which play fundamental roles in the way these molecules carry out their functions. To obtain a sufficiently precise atomistic description of a given molecular system it is necessary to describe its solvation accurately enough. There are two main different approaches to model the effects of aqueous environments in AMBER, implicit and explicit solvent.

Implicit solvent

Instead of explicitly simulating water molecules around the system, AMBER offers an implicit solvent alternative.

When estimating the solvation free energy of a molecule (ΔG_{solv}) we can decompose its calculation into the "electrostatic" (ΔG_{elec}) and "non-electrostatic" ($\Delta G_{nonelec}$) terms:

$$\Delta G_{solv} = \Delta G_{elec} + \Delta G_{nonelec} \tag{1.11}$$

where $\Delta G_{nonelec}$ is the free energy of solvating the molecule when all its charges are ignored and ΔG_{elec} is the difference between the solvation free energy of the molecule when all its charges are ignored and when they are added back.

$\Delta G_{nonelec}$ can be decomposed into two major opposed components: the favorable van der Waals interactions formed with the water molecules and the unfavorable breaking of the water-water interactions. In the current implementation of the implicit solvent in AMBER, $\Delta G_{nonelec}$ is proportional to the total surface accessible area of the molecule with a proportionality constant parameterized from experimental solvation energies of small non-polar molecules, being a fast term to calculate.

ΔG_{elec} has traditionally been solved by the Poisson Boltzmann equation (PBE) which is too computationally expensive to be used in molecular dynamics. A frequent solution is to use approximations to PBE, such as the analytic Generalized Born method implemented by AMBER developers.

Despite being considered less accurate[54,55], implicit solvent is widely used mainly for two reasons. First, it can significantly increase computational performance in terms of simulation time generated per processor time in comparison with an explicit solvent model, in which the interactions between all the solvent molecules have to be computed. Second, it can perform conformational sampling faster than explicit solvent[56–59] over the same period of simulated time. This effect can be expected to be derived from two main effects: the reduction of solvent viscosity and possible alterations on the potential landscape as interactions with solvent molecules are not explicitly modeled[60]. Therefore implicit solvent may be attractive to tackle conformational sampling problems. On the contrary, the resulting unreliability of the energy landscape and, most of all, its inability to explicitly represent water molecules that could play major roles in the protein function (e.g. catalytic roles) make this model completely unsuitable for the purposes of the simulations generated during this thesis.

Explicit solvent

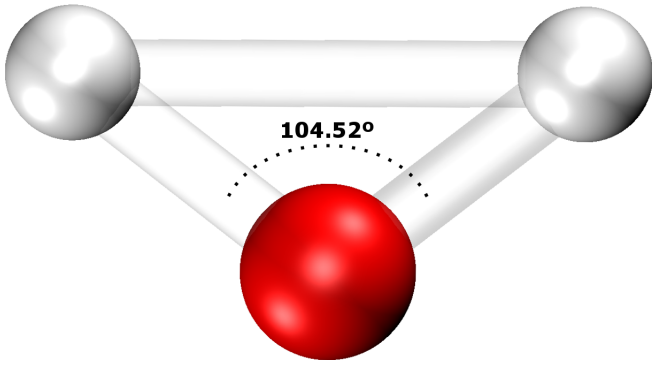


Figure 3: *Graphic depiction of a TIP3P water model molecule.*

In the explicit solvent approach a number of water model molecules, which are placed surrounding the system, are atomistically simulated. When using explicit solvent molecules, these usually become the major contributor to the number of atoms in the simulation and, therefore, a critical aspect regarding computational efficiency. A wide range of water models with varying levels of computational cost and accuracy exists[61–71]. In solvent focused simulations, in which solvent properties has to be reproduced as rigorously as possible, the usage of more expensive and accurate solvent models can be convenient. However, when working with biomolecules, the objective is usually to achieve a balance between a precise enough representation of the conformational energy landscape and a computational efficiency capable of reaching meaningful simulation times. Therefore classical force field for biomolecules, like AMBER[45,46] or CHARMM[72,73] are parameterized to use the simple but fast TIP3P water model. TIP3P is a 3-site water model in which, in addition to the typical O-H bonds in water, the molecule is rigidified by an internal H-H bond which keeps the H-O-H angle value at 104.52º (Fig. 3) and is parameterized to be used in conjunction with the SHAKE algorithm. SHAKE algorithm, in its most common use in AMBER, constrains all the bonds involving hydrogen atoms, removing hydrogen vibrations from the calculation. As the time step in a MD simulation has to be at least as short as the highest frequency motion in the system (i.e. hydrogen vibration), removing hydrogen vibration with SHAKE allows time steps of 2 fs instead of

the regular 1 fs, dramatically improving the computational efficiency of the simulation. This constraint also makes TIP3P molecules to be treated as rigid bodies because their three bonds involve hydrogen, which results convenient to further improve performance since TIP3P water model only has defined van der Waals radius for the oxygen atom, being both hydrogen atoms contained within this radius. The forced rigidity induced by SHAKE makes it impossible for hydrogen atoms to cross the van der Waals radius of the oxygen atom, eliminating a possible source of error due to this reduction.

Periodic boundary conditions

To avoid errors related with boundary effects, explicit solvent is commonly used under periodic boundary conditions (PBC) that establish a boundary box defining the primary cell. PBC treats the system as if it had an infinite extent, i.e. it was composed of an infinite number of copies of the primary cell in all directions. AMBER only keeps real track of one single set of atoms but calculates the forces related to the non-bonding term for all the atoms of the "infinite system" via imaging techniques (Fig. 4), actually only taking into account the relevant ones for the forces calculation (i.e. those within the non-bonding cut-off).

Once the initial structure has been thoroughly evaluated, to solvate it, LEaP module adds boxes of preequilibrated solvent model molecules until the distance from the box boundaries to any atom of the solute are at least as long as the given cut-off (12 Å in our case). In AMBER the shape of the PBC can be either a truncated octahedron or a cuboid box, being cuboid the most common shape and the one used in this thesis.

When using Ewald summation methods (e.g. PMEMD in AMBER) to simulate heterogeneous systems, such as proteins solvated in water, strong artifacts in the chemical potential of charged particles may occur if this systems have non-zero charge[74]. To avoid this, LEaP can calculate system net charge and add a given number of monoatomic counterions (e.g. Na+, Cl-) to neutralize it. After the solvation process has successfully finished, LEaP adds counterions around the solvent, replacing solvent molecules, using a Coulombic potential on a grid as the criterion to define the molecules to be substituted.

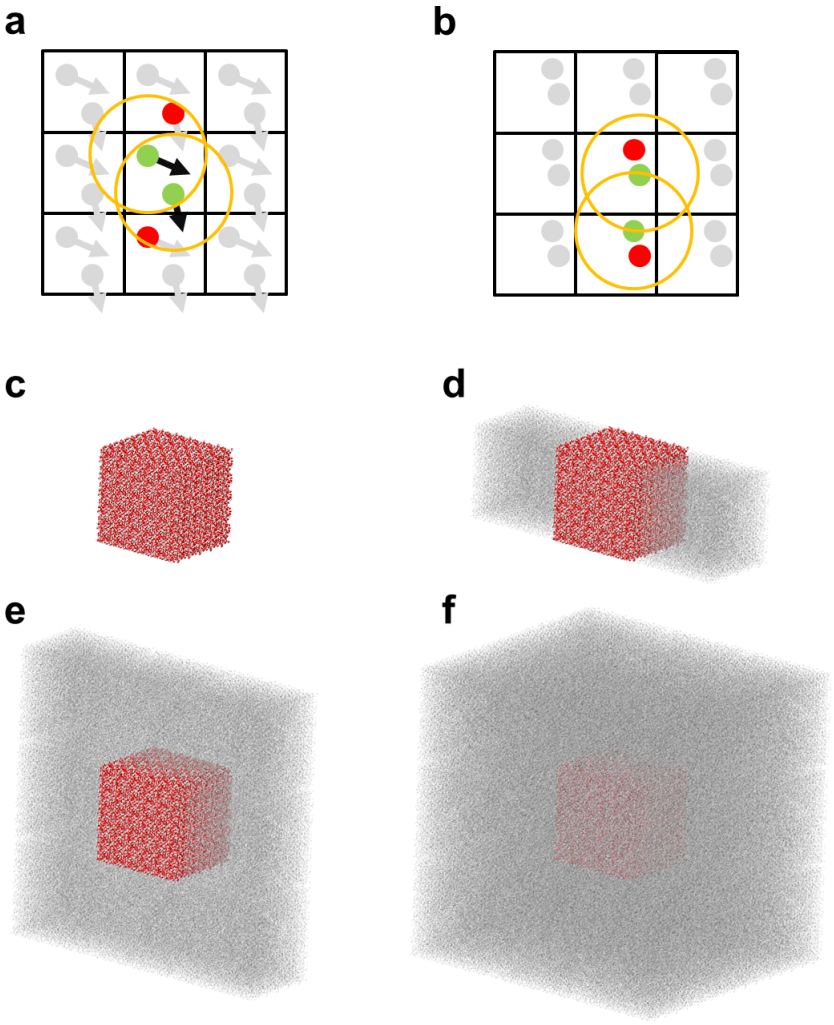


Figure 4: *Periodic boundary conditions. Figures* ***a*** *and* ***b*** *provide an schematic representation of two frames of a 2D simulation with a periodic boundary setup. Green circles represent the particles being simulated, gray circles represent the images generated by the boundary conditions and the orange circles depict the non-bonding cutoff radius, highlighting in red the images that fall inside the cutoff and, thus, those whose interaction with the green simulated particles is calculated. Note the high impact a change in the non-bonding cutoff is expected to have performance-wise. To offer a more realistic representation of the scale on which this principle was applied in the simulations present in this thesis, periodic images of a solvation box (***c***) are sequentially added along the x-axis (***d***), y-axis (***e***) and z-axis (***d***). White dots represent hydrogen atoms and red dots oxygen atoms. The solvation box surrounds protein that cannot be seen for clarity purposes.*

Input files

In a standard MD simulation with AMBER there are mainly three kinds of mandatory input files: "mdin" file (a file containing all the options and flags that define the simulation protocol), parameter-topology file and coordinates file.

Parameter-Topology file, also named "prmtop" file, essentially describes information that doesn't change during the simulation and, therefore, the same file is used for the whole simulation time, no matter how long it is. As its name indicates, contains two major sets of data: parameters and topology. Parameters refer to the charge, atomic number, mass, atom types, residue identifiers, and the molecular force field parameters for each type of atom, bond, angle and dihedral angles in the system. On the other hand, topology describes the connectivity of all the atoms, which is used to infer the bonds, angles and dihedral angles to be evaluated during each step. Additionally, if PBC are used, it lists the last residue to be considered solute, the first residue to be considered solvent, the total number of molecules and the total number of atoms in each molecule.

A coordinate file, also named "coord" file, in contrast, describes the information that changes during the simulation. Atomic coordinates are always present. If, during a MD simulation, the options specify that certain frame is to be saved with its velocities, they are written after the atomic coordinates, being the resulting file called "restart" instead of "coord". Lastly, if an MD is run under constant pressure or constant volume conditions, the size of the periodic box will also be added to the end of the file.

A "restart" file can be used to initiate a MD simulation using the velocities it contains, being a useful tool to create periodic backups during a simulation execution. This prevents major data losses in case the simulation is interrupted and also allows forking from the initial trajectory (e.g. during energy surface generation; see "Free energy surfaces" subsection).

Restraints

In addition to the harmonic potentials used in the MM force field to define bonds, angles and torsions, sometimes it is convenient to define ad-hoc potentials to modify the behavior of certain groups of atoms, e.g. to protect gaps in the structure from opening, to perform umbrella sampling, steered MD, etc. The most common way to generate these potentials in AMBER is via a restraint file, in which the parameters defining each potential will be specified (Fig. 5). In section 3.2.4) the restrains applied during the simulations presented in this thesis are detailed.

Minimization phase

Once the initial "coord" and "prmtop" files have been generated, three steps precede the generation of productive free MD trajectories.

First, the structure is subjected to a number of energy minimization steps (in our case 15000). A minimization process is a numerical method which pretends to find a particular arrangement of an ensemble of independent variables that corresponds to a minima in a

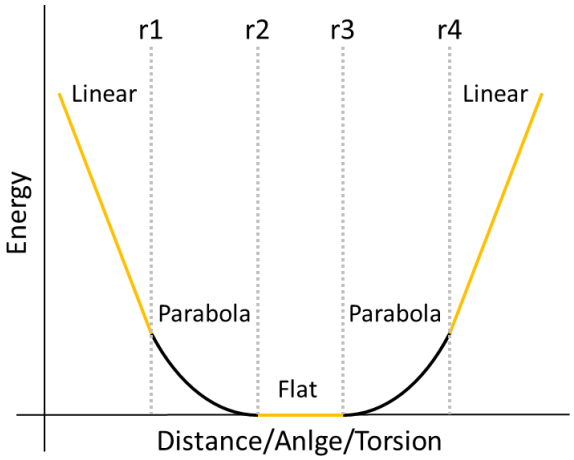


Figure 5: *AMBER restraint potentials. In AMBER, restraint potentials are defined using four measurement values r1, r2, r3 and r4, defining 5 regions for the potential. The region between r2 and r3 defines a flat-bottomed potential so that, if r2 and r3 differ, no energy penalty is applied when the system samples that region. On the other hand the potential shape of the other four regions is either a parabola or a linear function with independently configurable force constants (rk1, rk2, rk3 and rk4) so that the potential has not to be necessarily symmetric if the user does not want to.*

given energy function, starting from an initial configuration of those variables. In this case, in which a geometry optimization is the objective, the ensemble of variables is composed by the $3N$ atomic coordinates (being N the number of atoms) and the energy function the MM force field used.Given the extensive number of atoms present in biomolecular systems, little conformational space sampling can be achieved. Consequently only small rearrangements will be sampled and, therefore, convergence is extremely unlikely to happen. Nevertheless, this process is still worthy as it is highly effective solving the most problematic arrangements (e.g. collisions, abnormally long or short bond distances, etc.) that could have resulted from the modeling procedure, ligand placement or any other structure manipulation occurred during the initial structure preparation. In this thesis, the minimization protocol was performed with AMBER, applying the default combination of methods (steepest descent and conjugate gradient). During minimization and, most of all, equilibration phases the dihedrals of the Cα trace are restrained to prevent artificial conformational changes. Other structural parameters, such as ligand binding contacts, might also be restrained during this phase to prevent artificial disruptions that may induce possible artifacts.

Equilibration phase

After minimization, the structure is ready to be gradually heated to the working temperature (298 K). To introduce temperature in a system with no previous velocities, as is the case, AMBER generates initial velocities from a Maxwell-Boltzmann distribution based on the specified temperature value, and randomly assigns these velocities to the atoms in the system. Initial velocities are randomly assigned using a low initial temperature value (100 K in our case) and then the system temperature is raised at a constant rate (10 K per ps in our case) until the specified value is reached (298 K in our case). This process is divided in ten steps after each of which velocities are reassigned.

Stabilization phase

Once the thermostat is set to the final working temperature, a 20 ps long stabilization phase begins, which consists on progressively removing the restraints applied during the equilibration phase by gradually reducing the restraints force constants from their initial values to 0. Depending on the study to be performed, it might be convenient to define restraints that will be kept during the productive free MD trajectory. If that is the case, these restraints should be present from the minimization phase. In the case of this thesis restraints were used to protect structure gaps and to keep water molecules in catalytic configurations to improve sampling.

Free Molecular Dynamics

After minimization, equilibration and stabilization phases, the system is ready to yield a productive free MD trajectory. A basic control tool of the system stability during the simulation is the root mean square deviation (RMSD) of the new calculated frames with respect to a reference structure (typically the initial structure). The equation to calculate it in each frame can be written as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i}^{Atoms}(d_i^2)} \tag{1.12}$$

where N is the number of atoms over which RMSD is being evaluated and d_i is the distance between atom i in the two structures. Generally, structures are aligned first, minimizing the atom distances, obtaining a consistent measurement of structural deviation. For example, an RMSD value lower or equal to the resolution with which the initial structure (or the structure used for modeling) was experimentally solved is a good initial indication about the stability of the system. Depending on the system that is being simulated many other indicators should be periodically checked as well, such as the evolution of the different energy components (total energy, kinetic energy, restraint energy, etc.) or the root mean square fluctuation (RMSF) of important regions. RMSF represents the positional standard deviation of individual atoms and is frequently calculated over the whole Cα trace, offering a description of the positional dispersion of residues.

Steered Molecular Dynamics

Although MD simulation time scales typically range from nanoseconds to microseconds, reaching even milliseconds with dedicated supercomputers[75], dissociation of macromolecular interactions frequently occurs at time scales of seconds or larger[76–83]. In the case of this thesis, we were interested in the study of the separation of the ATPase heads of Smc1A and Smc3 of cohesin head. To illustrate the complexity of the problem, a good example of an ATP hydrolysis facilitated dissociation of proteins with experimentally measured rates is actin depolymerization (Fig. 6) being 8 s^{-1} the fastest and 0.3 s^{-1} the slowest. Therefore, with current computational resources, free MD simulations are not suited to study this kind of processes.

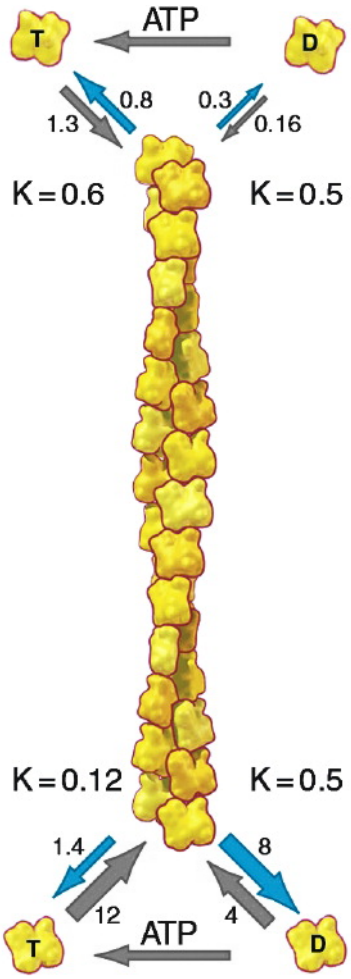


Figure 6: *Actin dissociation rate constans. Modified from Pollard and Borisy (2003)[76]. Arrows indicating dissociation processes have been kept in blue while the rest have been shaded in gray for clarity purposes.*

Steered molecular dynamics (SMD) is a biased MD technique, which facilitates the sampling of conformational space, otherwise unreachable over the current attainable simulation times, by applying an external potential to induce a change in a MD simulation. There are two kinds of SMD methods, constant velocity pulling and constant force pulling (Fig. 7). Constant force pulling method applies a constant force to one atom in the direction defined by the vector formed with a reference atom , which can be a dummy (an atom that do not participate in the rest of the calculations) or not. On the other hand, constant velocity method consist on applying a harmonic potential to alter a given coordinate, displacing the center of the potential in a specified direction at constant velocity. In a pulling simulation this is typically performed by adding a dummy atom which is attached to the atom to be pulled, moving the dummy atom in a given direction at constant velocity. In this thesis the method used is a modified version of constant velocity pulling in which the distance of the centers of mass of both subunits was controlled via a harmonic potential. The equilibrium distance was steadily increased over time, therefore obtaining a similar behavior to a constant velocity pull. The advantages of this modification are that no pull direction has to be established and no individual pull atoms have to be defined. Using the centers of mass to apply the pulling forces prevented conformational rearrangements close to the pull points that could have led to local artifacts.

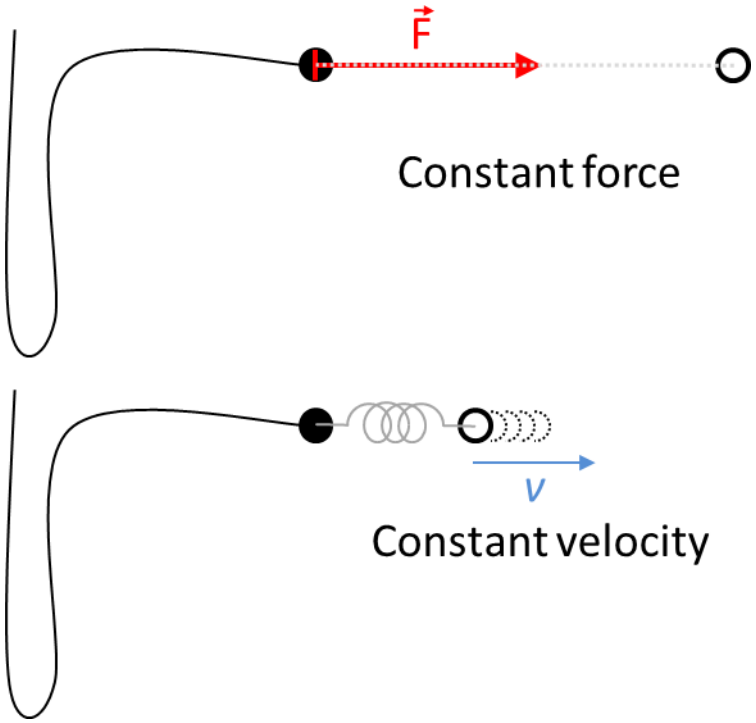


Figure 7: *Schematic representation of the constant force pulling and constant velocity pulling steered molecular dynamics methods.*

Jarzynski equality

To reconstruct free energy profiles (2D free energy surfaces) from the SMD trajectories of the ATPase heads of Smc1A and Smc3 Jarzynski equality[84] was used. Jarzynski equality states that the free energy difference between two quasi-equilibrium states is derivable from non-equilibrium transformations from one to the other through this equation:

$$\exp \frac{-\Delta G}{k_b T} = \left\langle \exp \frac{-W}{k_b T} \right\rangle \tag{1.13}$$

where ΔG is the free energy difference between states, k_b is the Boltzmann constant, T the temperature of the system, W is the accumulated work of each trajectory and bracket notation implies taking the average result of the calculations for each W. To perform Jarzynki calculations using the data generated with our center of mass SMD trajectories a universal implementation of Jarzynski was written in Python 3, capable of using data generated with any kind of distance-based potential. The code can be seen in appendix F.

Quantum Mechanics/Molecular Mechanics Molecular Dynamics

AMBER MM force field offers a potential energy function accurate and fast enough to study conformational dynamics and protein-protein interactions, which was the aim of the free MD and SMD studies performed in this thesis. However, to accurately describe chemical reactions and the effect the microenvironment has over catalysis, a quantum mechanical description is required. Unfortunately, the computational cost of quantum mechanics (QM) potentials is prohibitive in systems with so many atoms, even using geometry optimization methods from snapshots instead of full MD. The use of hybrid QM/MM (Quantum Mechanics/Molecular Mechanics) potentials is probably the most extended workaround to this problem. This approximation is based on the idea that QM considerations are only actually meaningful to describe the behavior of those atoms in close proximity to the chemical reaction. Consequently, in the QM/MM hybrid potential scheme the system is divided in two regions based on the potential energy function used to describe each: the QM region, and the MM region. The QM region is comprised of the atoms that could play an important role in the chemical reaction whereas the MM region contains the rest of the atoms (Fig. 8). Both regions interact with each other within a distance cutoff (Fig. 8), defining an interface region (QM/MM region). The criteria to define these regions will be discussed in the "Quantum Mechanics region definition" subsection. The MM function used in this thesis for QM/MM MD calculations was the same AMBER force field[45,46] used in the MD and SMD simulations. The QM potential was defined by Fireball[85,86]. The simulations were run with the Fireball/AMBER implementation[87,88] recently developed in collaboration between the group of Dr. José Ortega at the Department of Theoretical Condensed Matter Physics of the Autonomus University of Madrid and our laboratory.

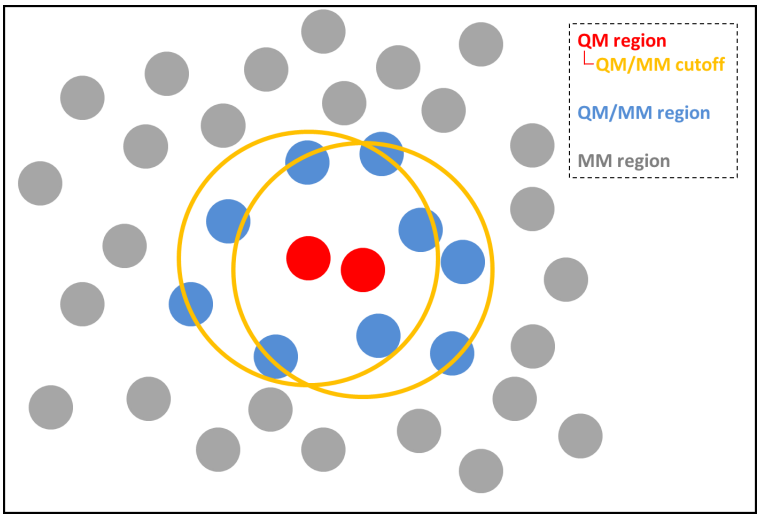


Figure 8: *2D scheme representing the interfaces present in a QM/MM MD simulation. Atoms simulated under QM conditions are depicted in red (QM region), atoms simulated under MM conditions whose non-bonding interaction with the QM region is considered in the calculations are shown in blue (QM/MM region) and atoms simulated under MM conditions whose interaction with the QM region is ignored are rendered in gray (MM region). The cutoff radius determining which atoms belong to the QM/MM region is indicated as an orange circle (QM/MM cutoff).*

Fireball

Fireball is a real-space local-pseudoatomic-orbital MD implementation of density functional theory (DFT)[86]. Here some of its characteristics will be introduced.

DFT is a computational method to approximate the quantum mechanical properties of a many-body system. It is based on the Born-Oppenheimer approximation which states that, although the forces acting on electrons and nuclei are of the same order of magnitude, nuclei are much more massive than electrons and, as a consequence, the kinetic energy of nuclei is much smaller than the kinetic energy of electrons. Based on this assumption, electrons are considered to instantaneously reach their ground state for a given nuclei configuration. Therefore, a ground-state potential energy function can be defined, over which nuclei are treated as classical particles. According to this notion, electronic states are calculated assuming that nuclei are stationary. In each state, the arrangement of these stationary nuclei induces a Coulomb potential that can be treated as a static external potential function of the nuclei positions. Therefore the system can be described as electrons interacting with each other and with an external potential (the nuclei Coulomb potential).

Modern DFT started with the Hohenberg-Kohn theorems[89], which demonstrated that, in a many-electron system under the effects of an external potential, 1) the ground and excited states properties can be exactly determined from the ground state electron density and 2) that this ground state electron density can be variationally calculated by minimizing an energy functional of the electron density, as the ground state energy is necessarily the minimum. In other words, given a nuclei configuration, if electron density can be determined, nuclei charge can be calculated and thus the corresponding Coulomb potential, which defines the external potential that is used to construct the electronic Hamiltonian. Unfortunately, although Hohenberg-Kohn second theorem demonstrates that the energy functional used to variationally calculate the ground state density can exist, it does not describe its form.

One year after the publication of the Hohenberg-Kohn theorems, Kohn and Sham[90] published a framework that made DFT tractable. In their approach they introduced orbitals and a fictitious system S of non-interacting electrons (i.e. electron-electron repulsion is ignored). The external potential in S is so that the ground state electron density of S and the one of the real system are the same. The advantage of using S is that the ground state wave function can be written in terms of simpler single-particle wave functions (known as Kohn-Sham orbitals), which, to ensure exchange anti-symetry (in this case that two electrons with parallel spins cannot occupy the same state), are introduced in a Slater determinant. We can decompose the total energy of the system E_{total} as a function of density ρ in:

$$E_{total}(\rho) = T_{elec}(\rho) + U_{Coulomb}(\rho) + E_X(\rho) + E_C(\rho) \tag{1.14}$$

where T_{elec} is the electronic kinetic energy, $U_{Coulomb}$ is the classic electrostatic energy, E_X is the exchange energy and E_C is the correlation (in this case correlation of electrons with anti-parallel spins) energy. Electron density is calculated from the Slater determinant and the spin orbitals in the determinant are used to obtain an estimation of the electronic kinetic energy T_s. Calculated electron density determines charge distribution and, thus,

classic Coulomb energy $U_{Coulomb}$. As T_{elec} has been calculated assuming non-interacting electrons, a correction T_{corr} has to be applied. The difference between $T_s + U_{Coulomb}$ and the real total energy necessarily is $T_{corr} + E_X + E_C$, which can be grouped in one single exchange-correlation term E_{XC} that is functional of the electron density.

There are many different approximations to calculate the E_{XC} term. In the Fireball implementation used to generate the simulations presented in this thesis, the BLYP[91,92] functional, a well-known GGA exchange correlation functional, is used, being further approximated by the McWEDA method[93].

The full calculation in Kohn-Sham DFT is done in a self-consistent manner, i.e. the orbitals are varied to minimize energy, which is itself calculated from those orbitals. In Fireball, Kohn-Sham DFT self-consistent functional is replaced by an approximate self-consistent functional based on atomic occupation numbers[94,95]. This approach uses computationally advantageous localized pseudo-atomic orbitals[93]. The interaction between these orbitals becomes zero beyond the cutoff radius, so integrals only have to be evaluated over a determined range. Additionally, approach integrals are pre-calculated once for each atom type, tabulated and the actual values are obtained by interpolation of those.

Fireball precision rests on using accurate basis sets for the set of atoms and conditions to be simulated. The Fireball basis sets for biological molecules have been developed on collaboration between the group of Dr. José Ortega and our laboratory, which has given rise to a PhD thesis[96]. One of those basis set, accounting for hydrogen, carbon, oxygen, nitrogen and phosphorus was the one used in the QM/MM simulations performed during this thesis.

Quantum Mechanics region definition

As previously introduced, in a QM/MM simulation the system is divided in QM and MM regions. QM treatment is a major performance bottleneck. Raising the number of atoms simulated under QM conditions can rapidly make the required calculation times unmanageable. On the other hand, it is mandatory to include all the atoms that could play any relevant role in catalysis in the QM region or, otherwise, severe artifacts could arise. Thus, the definition of a minimal, yet comprehensive, QM region is one of the most critical steps in QM/MM methods.

Our criteria to define the QM region were to introduce:

- All the atoms directly participating in the reaction and all their covalent neighbors up to non-polar bonds.
- All the atoms involved in direct electrostatic interactions with the atoms previously mentioned, also adding their covalent neighbors up to non-polar bonds when possible. In the case of interactions where the backbone polar atoms are participating, a QM region boundary can be introduced in a peptide bond.
- Any other atom with no direct interaction with the substrate/s that could be playing a role (e.g. proton transfer chains).
- It is usually advisable to restraint all the relevant interactions when the system

transitions from MM to QM/MM until the system becomes stable, with similar philosophy to the equilibration and stabilization phases of the MM MD protocol previously commented.

Once a QM region has been defined, it has to be thoroughly evaluated, paying special attention to the QM energy evolution and convergence. Therefore, defining a stable and comprehensive QM region usually is an iterative process in which it is redefined and reevaluated several times.

Transition State Theory on biomolecular systems

In this thesis QM/MM MD simulations were used to compare the reactivity of cohesin active sites in different conditions from a classical transition state theory (TST) perspective, a widely used theory of the rates of elementary reactions. The relevance of TST is partially derived from its simplicity of use and interpretation in comparison with other theories of rates[97]. The first major assumption TST does is to consider thermodynamic quasi-equilibrium between the reactants and transition state, being the transition state the highest energy point of the free energy profile of the reaction (Fig. 9). TST states that the rate constant of a reaction k is given by[98]:

$$k = k\frac{k_B T}{h} K^{\ddagger} \tag{1.15}$$

where k_B is the Boltzmann constant, T the temperature of the system, h the Planck constant, k the transmission coefficient and $K^{\ddagger}$ the equilibrium constant between reactants and transition state. This is built upon the second assumption made in TST, the premise that only a one way flux is considered. In other words, a trajectory that crosses the transition state never crosses it back as part of the same event. However, this does not imply that, in other time scale, once the product is thermalized it may undergo the reverse reaction as in fact, if given enough time, it will[99].

Rate constant can also be written as function of the entropy and enthalpy of activation[98]:

$$k = k\frac{k_B T}{h} \exp\big(\frac{\Delta^{\ddagger} S^{\circ}}{R}\big) \exp\big(-\frac{\Delta^{\ddagger} H^{\circ}}{RT}\big) \tag{1.16}$$

where R is the gas constant, $\Delta^{\ddagger} S^{\circ}$ is the entropy of activation (the difference between the entropy of the transition state and the reactants), $\Delta^{\ddagger} H^{\circ}$ is the enthalpy of activation (the difference between the enthalpy of the transition state and the reactants). This equation is also usually referred to as the Eyring equation[100,101]. It can be equivalently written as function of the Gibbs free energy of activation[98]:

$$k = k\frac{k_B T}{h} \exp\big(-\frac{\Delta^{\ddagger} G^{\circ}}{RT}\big) \tag{1.17}$$

where $\Delta^{\ddagger} G^{\circ}$ is the Gibbs free energy of activation (the difference between the Gibbs free energy of the transition state and the reactants; see Fig. 9 a and b). The transmission

coefficient k present in the equations allows the possibility of the transition state not giving rise to a particular set of products if more than one exists[98]. Thus, in reactions with a single product it is often ignored.

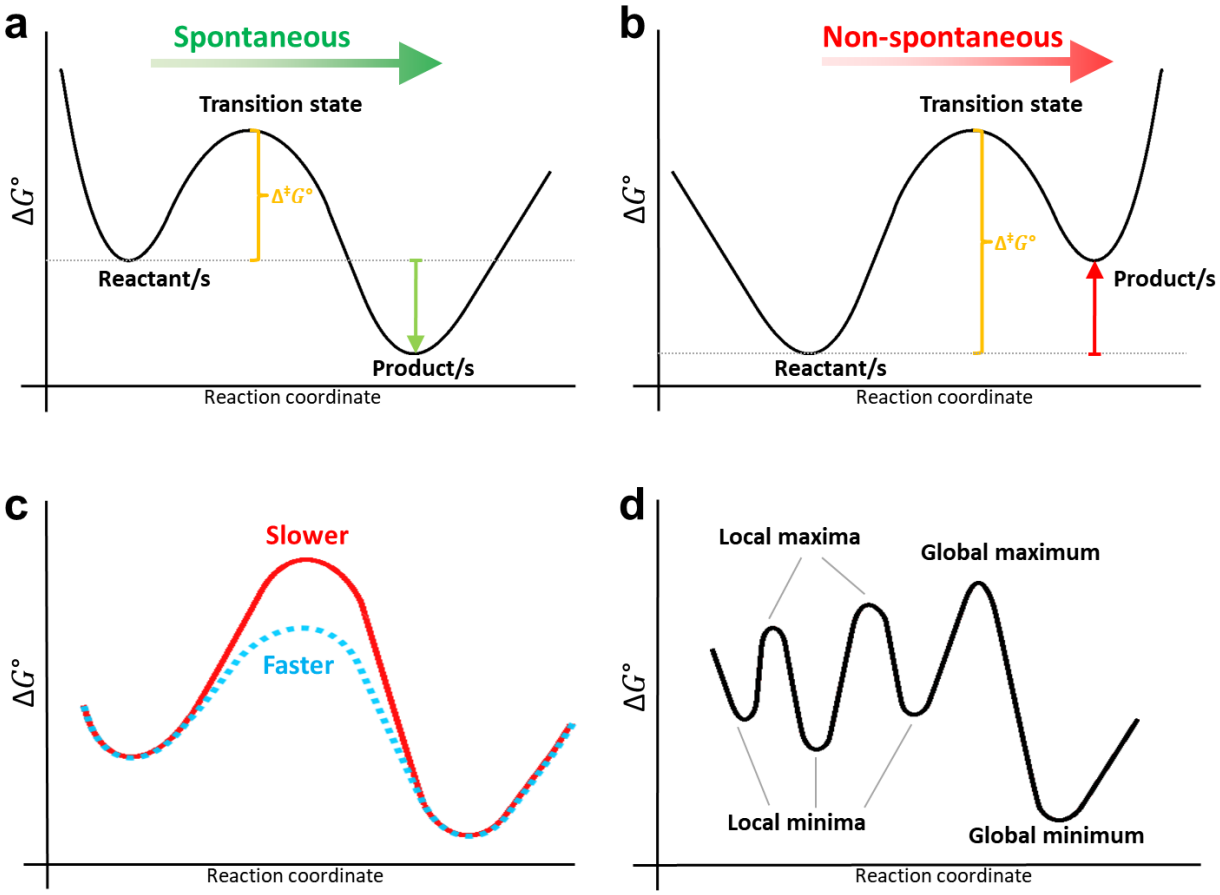


Figure 9: *Free energy profile interpretation in transition state theory. Free energy profiles describing spontaneous (**a**) and non-spontaneous (**b**) reactions are compared, indicating the free energy of activation in yellow and the free energy difference between reactant/s and product/s in green (spontaneous) and red (non-spontaneous) vertical arrows. In figure (**c**) two free energy profiles differing in the free energy of activation are presented, being the one with the lowest barrier (blue) the one describing the fastest transition between states. Figure (**d**) presents a free energy profile in which local and global maxima and minima are indicated.*

TST was used in two ways in this thesis: as a theoretical basis for the algorithms present in MEPSA and to estimate the difference in reactivity between conditions as, if we can estimate $\Delta^{\ddagger}G^{\circ}$ for two given conditions, we can approximate the difference in their reactivity.

Free energy surfaces

Calculating $\Delta^{\ddagger}G^{\circ}$ requires defining the geometry of the transition state and the corresponding difference between the Gibbs free energy of the reactants and that geometry. As Gibbs free energy is function of the atomic coordinates, to strictly evaluate all the possible states in a system with N atoms, a $3N-6$ dimensions free energy surface should be calculated. Not only this is computationally intractable but also the interpretation of

such a complex structure would be extraordinarily challenging. Instead, only a subset of coordinates capable of describing the relevant states of the undergoing process is used, letting the rest of coordinates relax after the perturbation, moment in which the system would exhibit a Maxwell-Boltzmann energy distribution. In this thesis 2D (one coordinate and energy) and 3D (two coordinates and energy) Gibbs free energy surfaces were calculated.

ATP hydrolysis consists on an oxygen atom of a water molecule attacking the γ-phosphate group of the ATP (bond to be formed), which leads to the disruption of the bond between γ-phosphate and β-phosphate groups (bond to be broken), the interaction of which becomes severely unfavorable when the electrostatic effects become dominant (both groups have negative net charges; see Fig. 10). Consequently, the reaction coordinates chosen to sample 3D free energy surfaces were these two bonds whereas the water molecule attack distance was the coordinate used to obtain 2D free energy profiles. Note that, for clarity purposes, in this text 2D free energy surfaces are termed 2D free energy profiles.

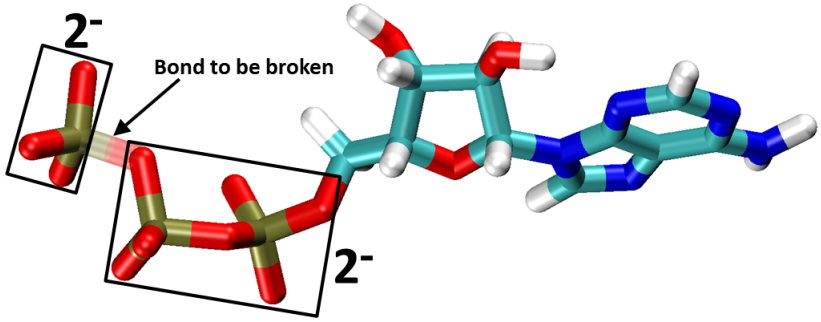


Figure 10: *Negative charges in ATP hydrolysis. The broken bond in a γ-phosphate group hydrolysis is indicated (bond to be broken). The leaving γ-phosphate group presents two net negative charges, while α-phosphate and β-phosphate groups present one negative charge each.*

To generate these surfaces, once the system was properly stabilized in QM/MM MD, the sampling along the reaction coordinates was performed via SMD by directly restraining those coordinates, yielding 7.6 x 10^6 structures for 3D surfaces and 7.7 x 10^4 for 2D profiles.

2D profiles were sampled by forcing the shortening of the distance describing the bond to be formed. Then, from the starting points generated during that trajectory along a single reaction coordinate, energies were sampled with 77 simulations with static restraint values distributed every 0.025 Å (Fig. 11). The last 10^3 energy values were used for the profile calculation.

3D surfaces were generated in two steps (Fig. 12):

- A SMD trajectory sampled along the reaction coordinate describing the bond to be broken, while keeping the attack distance fixed.
- Every 0.025 Å an initial structure was extracted from the first SMD trajectory. These structures were used to sample the reaction coordinate describing the attack distance while keeping the broken distance fixed in those 0.025 Å intervals.

Sampled energy values were distributed in groups along a grid. The energies in each bin of the grid were averaged, being weighted by the probability given by their partition

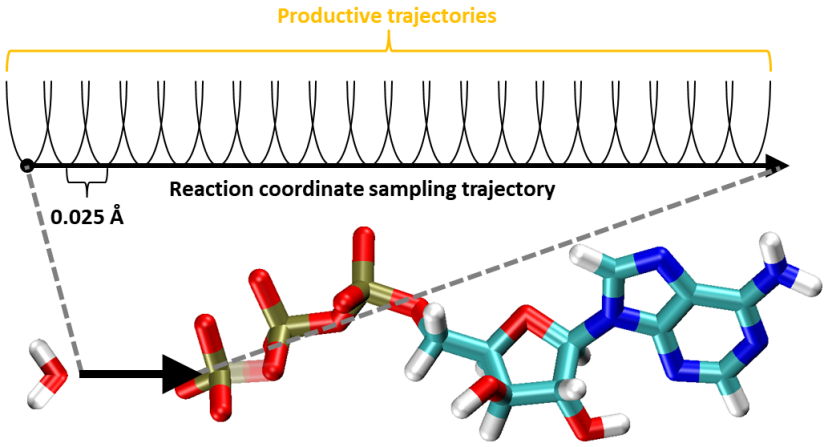


Figure 11: *Schematic representation of 2D profile generation protocol. The chosen reaction coordinate is sampled along an initial trajectory. Every 0.025 Å a productive trajectory is generated keeping the reaction coordinate restrained.*

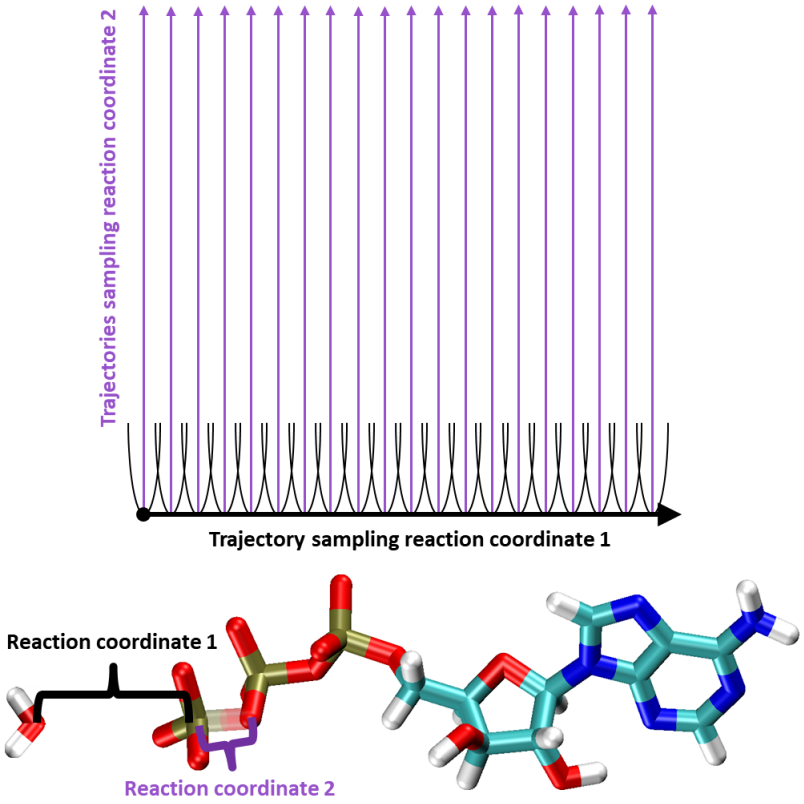


Figure 12: *Schematic representation of 3D surface generation protocol. One reaction coordinate is sampled along an initial trajectory in which the other reaction coordinate is restrained. Every 0.025 Å productive trajectories are generated, sampling the second reaction coordinate while keeping the first reaction coordinate restrained.*

function. Finally, a LOESS local regression smooth was applied.

From a classical TST perspective the key elements to observe in a free energy profile are local minima connected by local maxima (Fig. 9). Local minima represent the stable states of the system and local maxima the transition states connecting them. Local minima energy level determines the spontaneity of transitions while local maxima energy

level governs the transition rates between states. Given two local minima states A and B connected by a transition state T, B has to be a state with lower energy than A for $A \rightarrow B$ transition to be spontaneous (Fig. 9 a and b). On the other hand, the energy level of T regulates the transition rate between A and B, being the lower the barrier the faster the transition (Fig. 9 c). Local minima between A and B are usually referred intermediary states (Fig. 9 d) and the transition state with the highest energy value is considered the TST transition state of the reaction, as it represents the actual bottleneck of the process.

On contrast, 3D free energy surfaces contain more information and allow comparing different reaction pathways. The relevant elements in these are local minima connected to one another through saddle points (Fig. 13). Local minima still represent the stable states of the system but, now, transition states are defined by saddle points and local maxima just describe unlikely states (Fig. 14). If we evaluate the first partial derivatives of G° with respect to the two reaction coordinates x and y, a point of the surface is a critical point (i.e. minimum, maximum or saddle point) if:

$$\frac{\delta G^\circ}{\delta x} = 0 \tag{1.18}$$

$$\frac{\delta G^\circ}{\delta y} = 0 \tag{1.19}$$

A critical point is a saddle point if:

$$\frac{\delta^2 G^\circ}{\delta x^2}\frac{\delta^2 G^\circ}{\delta y^2} - \frac{\delta^2 G^\circ}{\delta x \delta y} < 0 \tag{1.20}$$

In other words, a saddle point is a point with gradient 0 in both directions (i.e. a critical point) which is a local maximum along one coordinate and a local minimum along the other (Fig. 13). The saddle point is the point of the barrier most likely to be visited, as it is the minimum along that barrier and, in a Boltzmann distribution, the probability of a state of energy E is:

$$P(E) = \frac{\exp(-\frac{E}{k_B T})}{Z} \tag{1.21}$$

where k_B is the Boltzmann constant, T the temperature of the system and Z is the partition function:

$$Z = \sum_{E_i} \exp(-\frac{E_i}{k_B T}) \tag{1.22}$$

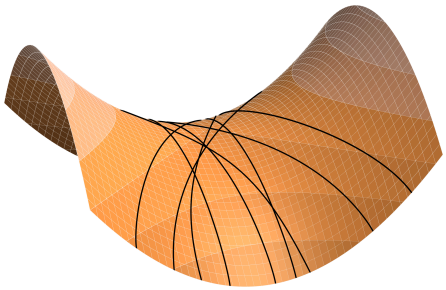


Figure 13: *Schematic representation of a saddle point crossed by a set of random trajectories.*

Consequently, although less favorable points of the barrier are also visited (Fig. 13), probability is narrowly centered on the saddle point. Given this negative exponential function, describing the minimum energy path connecting two states is highly descriptive of the transition kinetics, also simplifying the comparison between different paths (i.e. different catalytic mechanisms) on the same surface and/or on different surfaces. This was the aim of one of the results obtained during this thesis, the development of MEPSA, which was used to analyze 3D free energy surfaces to obtain the minimum energy paths between substrates and products (Fig. 15 a). These paths were used to obtaining the corresponding energy profile over which points of interest could be easily visualized (Fig. 15 b). This facilitated the assignment of representative geometries to each point of interest and allowed the generation of a trajectory describing the reaction along the minimum energy path (see section 3.2 "Two-step ATP-driven opening of cohesin head"; figure 51 and appendix C). A detailed description of MEPSA development and usage is available in section 3.1.

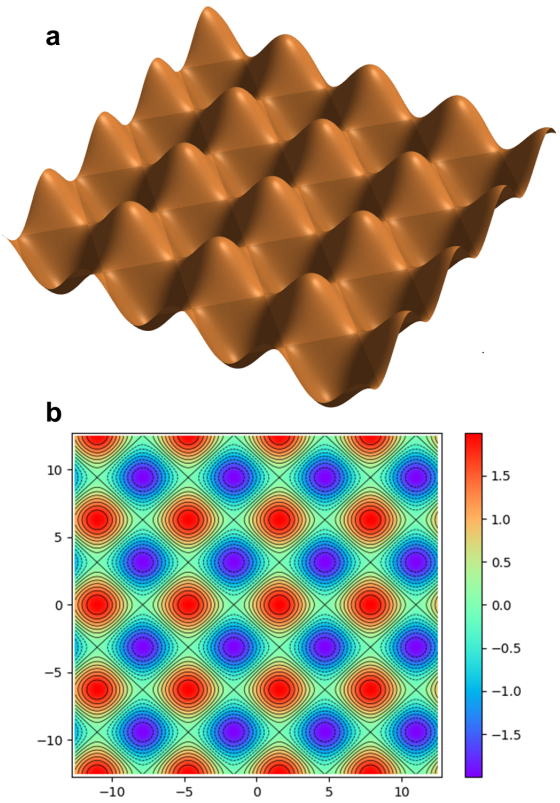


Figure 14: *3D (**a**) and contour (**b**) plots of a $sin(x) + cos(y)$ surface illustrating the concept of maxima and minima connected by saddle points in free energy surfaces.*

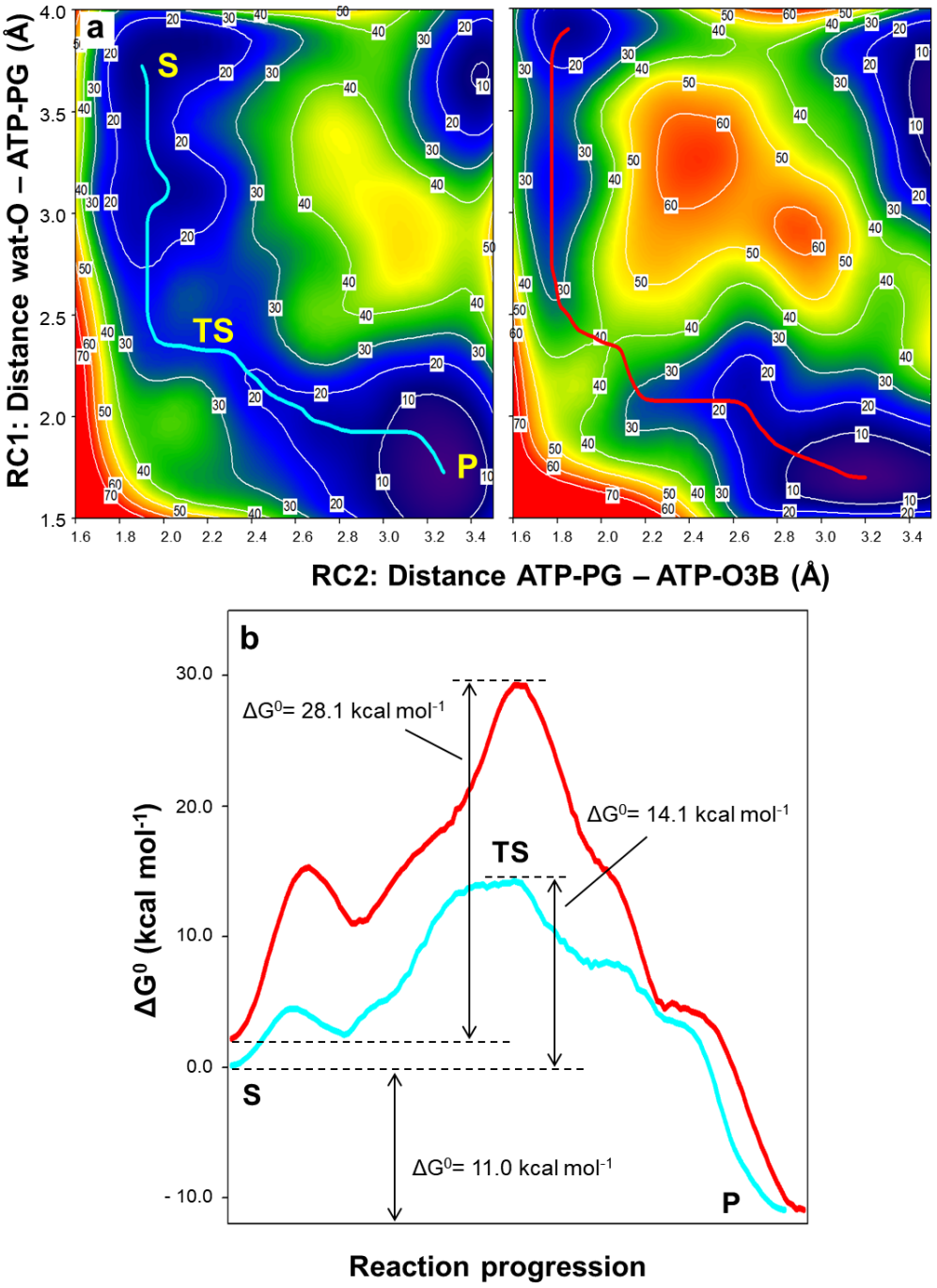


Figure 15: *Free energy surface analysis presented in Marcos-Alcalde et al. (2017)[24]. In section 3.2 "Two-step ATP-driven opening of cohesin head" details about the generation of these results and the subsequent interpretations are discussed.*

1.2.4 Molecular Docking

Automated molecular docking is a computational method to approximate the optimal binding configuration of two molecules offering a binding affinity estimate. Here we will focus on protein-ligand docking, as is the technique used in this thesis to infer possible druggable compounds that would bind to a pocket formed during the MD simulations. Molecular docking software uses search algorithms to sample configurations of the ligand which are generally evaluated with an empirical scoring function. The complexity of both the sampling algorithm and the scoring function is determinant to the computational performance. There are two typical uses of automated protein-ligand docking: i) to predict as accurately as possible the binding mode of one or few ligands and ii) to perform virtual screening, i.e. estimate the binding affinity of a large library of compounds, which usually requires faster methods. This was our case, as we wanted to evaluate a first set of 130374 molecules (the molecules available in the biogenic dataset from the ZINC15 database[102] in 2D format at that time) and larger ones in the future. The docking software used was smina[103], a fork of AutoDock Vina[104] (that will be referred to as Vina in this text) under active development which supports the implementation of custom scoring functions that could be useful in the future. However, the way smina was used to obtain the docking results that are shown in this thesis makes it indistinguishable from Vina and all the operation details explained bellow are equally descriptive for both programs. The binding energy predicted by the Vina scoring function has two parts, one depends on the conformation and the other on the number of active rotatable bonds. The conformation-dependent part is a functional with the following form:

$$c = \sum_{i<j} f_{t_i t_j}(r_{ij}) \tag{1.23}$$

Where, for a pair of atoms i and j, interaction functions f are given by the atom type of i and j and are functions of the interatomic distance r_{ij}.

The interaction energy functions $f_{t_i t_j}$ chosen by Vina developers are a mixture of knowledge-based potentials and empirical scoring functions[104] and can be decomposed in 5 terms:

$$f_{t_i t_j} = \begin{cases} w_1 Gauss_1(d_{ij}) + \\ w_2 Gauss_2(d_{ij}) + \\ w_3 Repulsion(d_{ij}) + \\ w_4 Hydrophobic(d_{ij}) + \\ w_5 HBond(d_{ij}) \end{cases} \tag{1.24}$$

where w_{1-5} are the weights for each term (Table 1) and d_{ij} the surface distance, which is given by:

$$d_{ij} = r_{ij} - R_{t_i} - R_{t_j} \tag{1.25}$$

where R_{t_i} and R_{t_j} are the van der Waals radii for each atom type t.

Weight	Term
−0.0356	gauss1
−0.00516	gauss2
0.840	repulsion
−0.0351	hydrophobic
−0.587	hydrogen bonding
0.0585	Nrot

Table 1: *Weights for each term in Autodock Vina scoring function (Source: Trott and Olson (2010)[104]).*

Steric interactions are modeled by:

$$Gauss_1(d_{ij}) = e^{-(d_{ij}/0.5)^2} \tag{1.26}$$

$$Gauss_2(d_{ij}) = e^{-((d_{ij}-3)/2)^2} \tag{1.27}$$

$$Repulsion(d_{ij}) = \begin{cases} d^2, \text{ if } d < 0 \\ 0, \text{ if } d \geq 0 \end{cases} \tag{1.28}$$

Hydrophobic interactions are modeled with $Hydrophobic(d_{ij})$, which equals 1 when $d_{ij} < 0.5$ and 0 when $d_{ij} > 1.5$. For d_{ij} values in the range $0.5 > d_{ij} < 1.5 Hydrophobic(d_{ij})$ is linearly interpolated between 0.5 and 1.5.

Similarly, the energetic contribution of hydrogen bonds is described by $HBonds(d_{ij})$ which equals 1 when $d_{ij} < -0.7$, 0 when $d_{ij} > 0$ and, for d_{ij} values in the range $-0.7 > d_{ij} < 0$, it is linearly interpolated between -0.7 and 0.

The energy contributions to the conformation-dependent part of the Vina scoring function can be decomposed in intramolecular (c_intra) and intermolecular (c_inter) terms:

$$c = c_{intra} + c_{inter} \tag{1.29}$$

The optimization algorithm, based on the Iterated Local Search global optimizer[105,106], samples conformations using c as criterion. On each step a random movement is introduced, the new conformation is locally optimized and the result is evaluated under a Metropolis criterion that determines if it is accepted or rejected[104]. Through this procedure the algorithm yields a series of conformations that are ranked according to their c score. The lowest-scoring conformation is then used to estimate the binding free energy with the conformation independent function g:

$$g(c_{inter1}) = c_{inter1}/(1 + wN_{rot}) \tag{1.30}$$

where c_{inter1} is the intermolecular term of c for the lowest-scoring conformation, N_rot is the number of active rotatable bonds and w is the weight given to N_rot (Table 1). This is the energy value Vina returns to the user.

Although in Vina the time spent on the search algorithm is varied heuristically depending on various parameters (number of atoms, flexibility, etc.), the search time can be increased

or decreased via the exhaustiveness parameter. Computation time scales linearly with exhaustiveness but the probability of missing the minimum energy configuration decreases exponentially. Additionally, in smina many other parameters can be tuned.

In the case of this thesis, to reduce the screening computation time, we performed three consecutive docking runs using increasingly detailed sampling together with progressively restrictive binding energy cutoffs. During the two first phases maximum minimization steps were set to 5 and exhaustiveness to 2 and 6 respectively. In the third phase minimization was allowed to be automatically scaled and exhaustiveness was set to its default value (8). The benefit of using this strategy was that extremely unfavorable molecules, most of which were also large and therefore computationally expensive to sample, were removed in the two first less expensive phases, significantly accelerating the process. The parameters used in each phase were conservatively chosen from previous tests in which such parameters yielded undetectable rates of false negatives.

The library of compounds used in this thesis was downloaded from ZINC15[102]. ZINC15 is a database of over 400 million small molecules (started with over 120 on release) which offers a user-friendly interface to select subsets based on a wide range of criteria (degrees of commercial availability, molecular size, reactivity, charge at different pH values, synthetic or biogenic origin, drug approval, etc.).

The docking results presented in this thesis were a proof-of-concept and are part of a work in progress, in which we aim to optimize the pipeline and work with larger sets of ligands in the future.

2 | Thesis objectives

The objectives of this thesis were:

- To generate an atomistic dynamic model of cohesin complex ATPase head heterodimer as detailed as possible, in order to tackle open questions regarding the role and mechanisms of ATP hydrolysis in the context of the cohesin function.
- To develop tools to efficiently interpret the resulting data, particularly focusing on standardizing the analysis of free energy surfaces, which has become the standard workhorse in the group to study enzymatic reactions.
- To use the resulting cohesin complex head heterodimer models to:
  - Gain mechanistic understanding of the role that ATP hydrolysis could be playing in the human cohesin head heterodimer.
  - Identify residues that could be playing key previously unreported roles.
  - Rationalize pathogenic variants.
  - Rationalize mutant phenotypes which are not well understood yet.
  - Generate a framework for *in silico* drug discovery.

3 Results

In this section, the main results related with the exposed objectives are detailed. Results derived from the generation, interpretation and use of an atomistic model describing cohesin ATPase dynamics can be found in sub-sections 3.2 "Two-step ATP-driven opening of cohesin head" and 3.3 "Allosteric coupling inhibitor screening via molecular docking". In addition, a set of in-house tools to facilitate the study of enzymatic reactions, such as the ATPase activity of cohesin, was developed, streamlining the analysis of 2D free energy surfaces to estimate and compare the activation free energy and the free energy difference between substrate and product, as well as obtaining structural descriptions of each of the relevant critical points. Eventually, the functionalities of this set of tools were gradually extended, a graphical user interface was developed and it was made publicly available as MEPSA, being also published in a peer-reviewed journal[23]. The description of MEPSA and a practical example of its use are presented in section 3.1 "MEPSA: minimum energy pathway analysis for energy landscapes".

3.1 MEPSA: minimum energy pathway analysis for energy landscapes

3.1.1 Introduction

The dramatic increase in computational performance along with the development of efficient sampling methodologies is making the calculation of energy landscapes defined by two reaction coordinates more accessible over time. These 3D energy surfaces can be used to study a wide range of molecular phenomena (nucleotide or protein folding[107,108], ligand binding[109], enzymatic reactions[88], etc) but the extensive amounts of information they offer might sometimes be laborious to process.

Since our group started calculating 3D free energy surfaces from QM/MM MD calculations the protocols for both generation and analysis of free energy surfaces were continuously tuned and improved in the light of the acquired experience. The first and simpler surfaces calculated exhibited single transition states[110,111] and required a straightforward three point analysis which could be tackled with little effort. During the next works[87,88], due to the improved capabilities of Fireball/AMBER in comparison to the semiempirical methods previously used, surfaces increased in complexity and we started to develop an internal set of analysis tools to streamline the analysis of the minimum energy path from substrate to product that was successfully used in those works (Figures 16 and 17). Several new functionalities were gradually added to the program, which at that point had become the standard tool for surface analysis in the lab. When the code consolidated we decided to publish program, MEPSA (Minimum Energy Path Surface Analysis)[23], under an open source GPLv3 license, so that it could be useful for other groups in the future. After its publication, MEPSA continued to be used in subsequent works[24,112] and other authors[113,114] (Figures 18, 15, 19 and 20), making use of other features such maxima edge profile detection (Fig. 18) or the SMD restraints generation from a calculated minimum energy path (appendix C). MEPSA[23] is a python program capable of performing several analyses from a transition state theory point of view in a user-friendly fashion due to its graphical user interface (GUI). There is no restriction related to the technique used to generate the surface or to its complexity as long as it is rectangular and uniformly distributed.

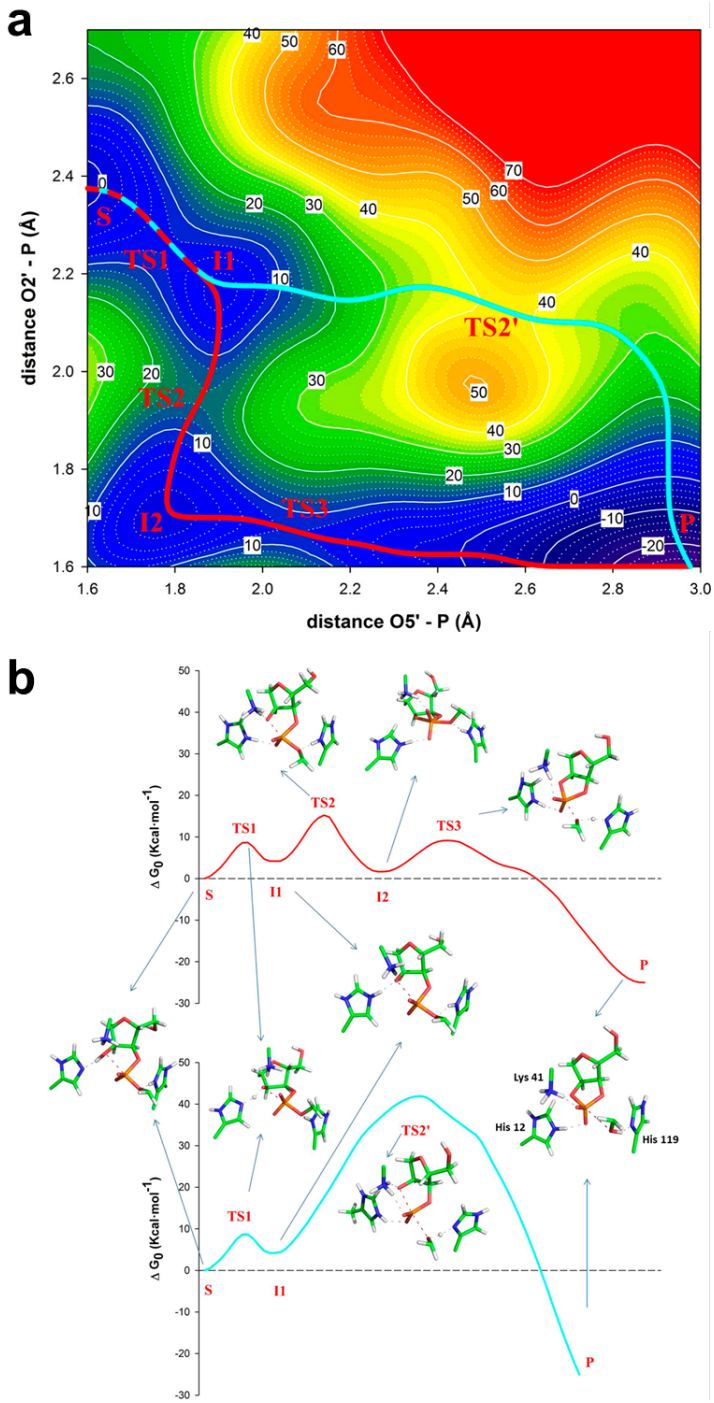


Figure 16: *First published results processed with MEPSA before public release (Source: Mendieta-Moreno et al. (2014)[87]). MEPSA was used to compare two possible pathways (dissociative S_N1 in cyan and associative S_N2 in red) for RNA cleavage by RNAase A (**a**) and to obtain representative structures of the key steps revealed by the free energy profile of both pathways (**b**).*

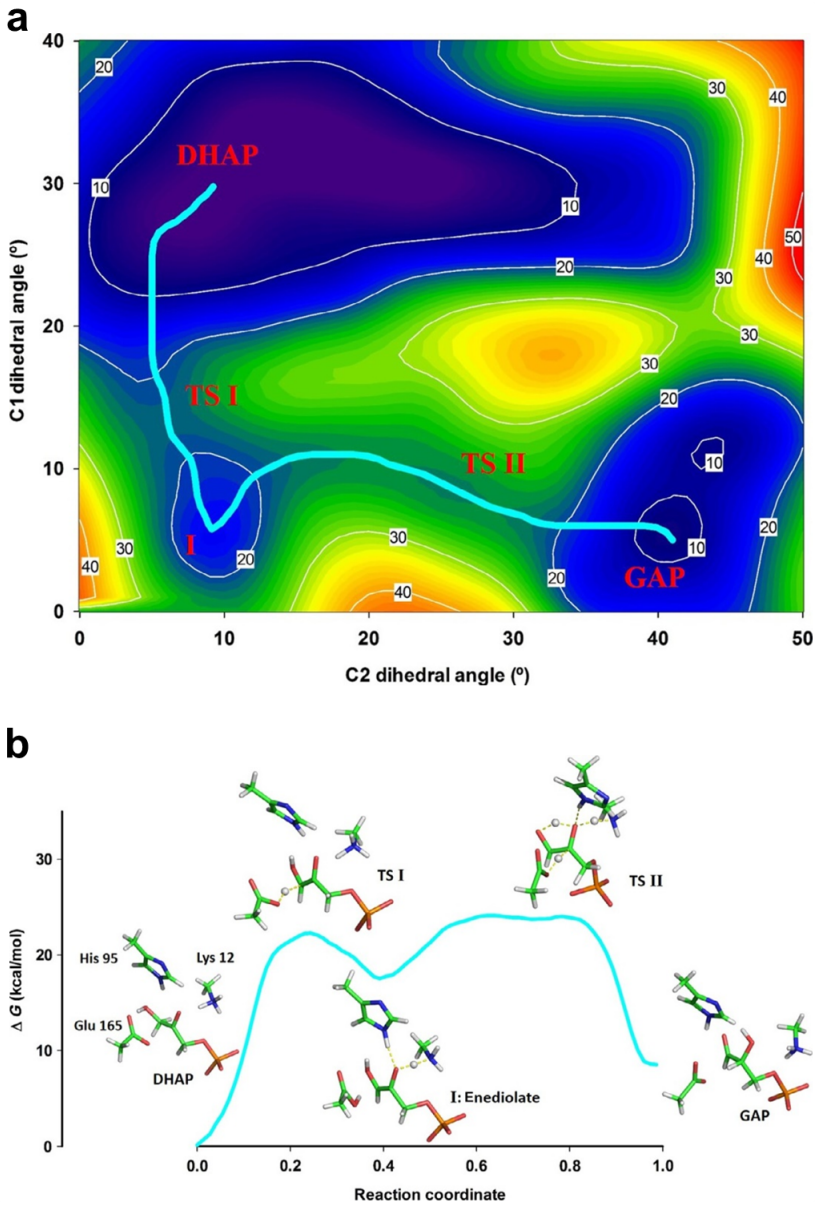


Figure 17: *Second published results processed with MEPSA before public release (Source: Mendieta-Moreno et al. (2015)[88]). MEPSA was used to determine the minimum energy path describing the isomerization mechanism catalyzed by the triose-phosphate isomerase (**a**) as well as to obtain representative structures of the key steps revealed by the free energy profile of the predicted pathway (**b**).*

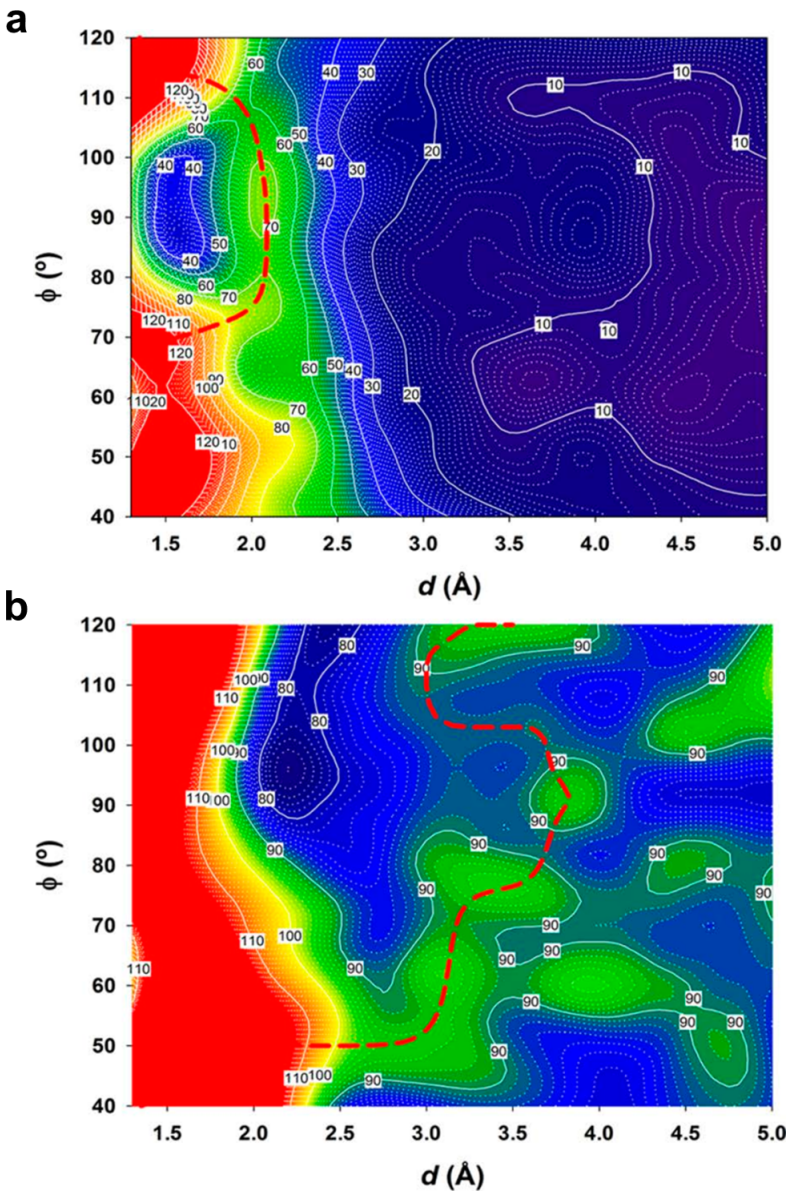


Figure 18: *First published results processed with MEPSA after public release (Source: Mendieta-Moreno et al. (2016)[112]). MEPSA maxima edge profiling functionality was used to compare the free energy barriers for cyclobutane thymine dimer formation between the ground (**a**) and excited (**b**) states of a pair of adjacent thymine nucleosides.*

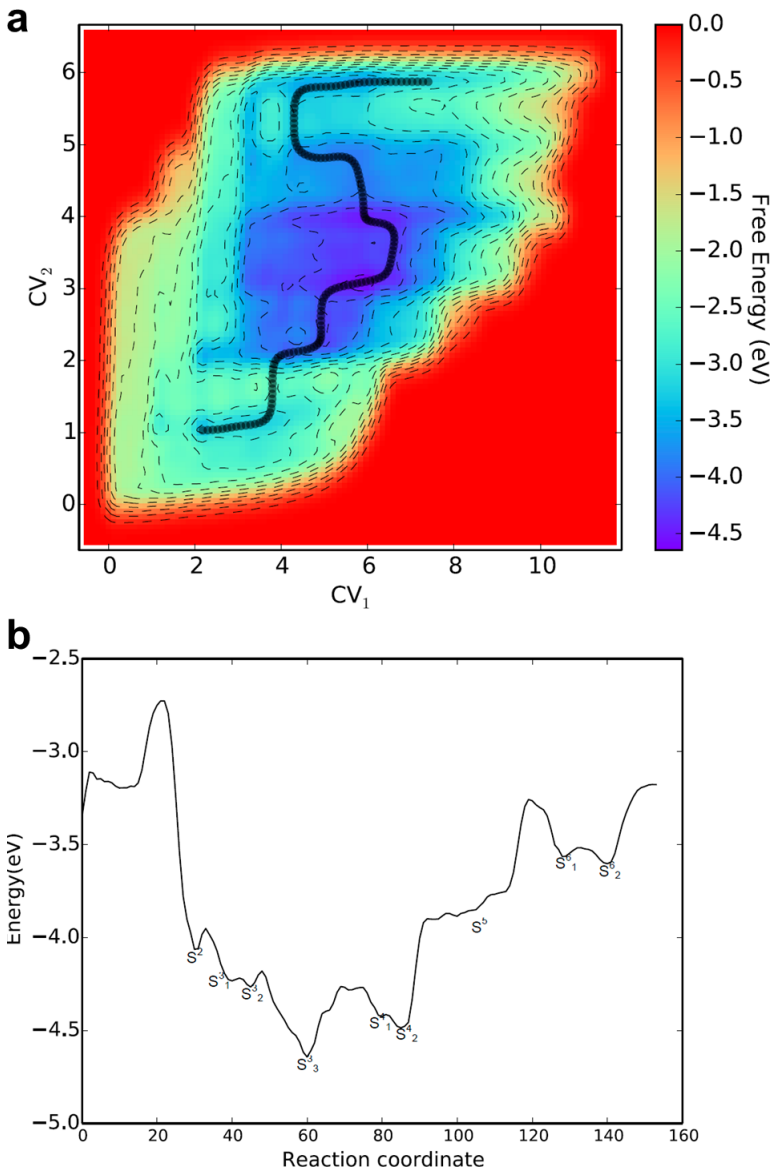


Figure 19: *Third published results processed with MEPSA after public release (Source: Kachmar et al (2017)[113]). MEPSA was used to determine the most favorable conformations of lithium solvation in ethylammonium nitrate. Results based on the energy profile (**b**) of the calculated minimum energy path (**a**) showed that coordination with 3 (S^3 states) or 4 (S^4 states) nitrate molecules was the most favorable.*

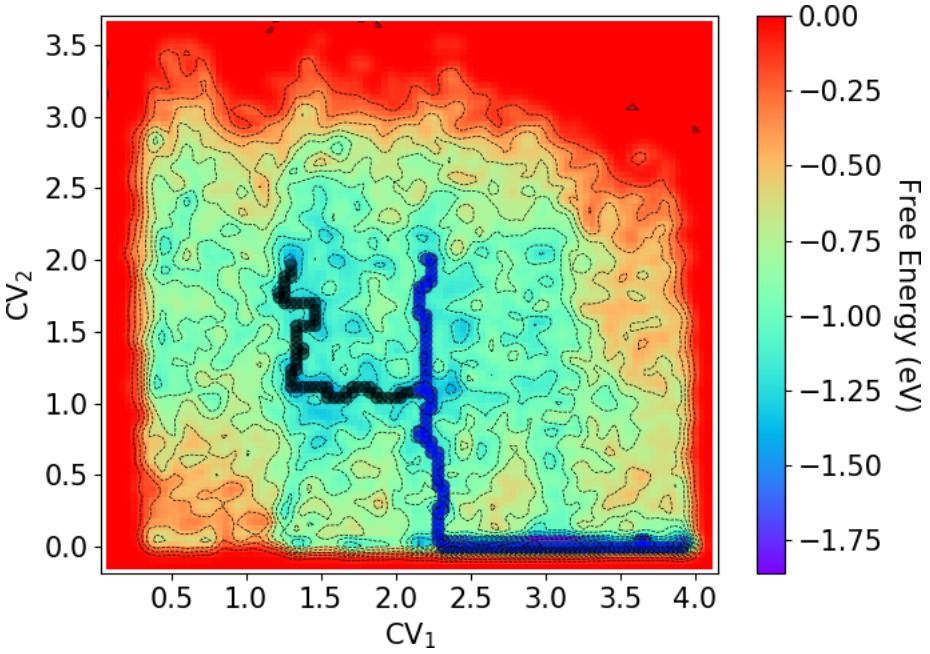


Figure 20: *Fourth published results processed with MEPSA after public release (Source: Kachmar et al (2018)[114]). MEPSA was used to determine the most favorable conformations of sodium solvation in dimethyl sulfoxide. Results based on the energy profile (**b**) of the calculated minimum energy path (**a**) showed that coordination with 3 or 4 dimethyl sulfoxide molecules was the most favorable.*

3.1.2 Availability

MEPSA is open source software (under a GPLv3 license) and can be freely downloaded from the CMBSO Molecular Modeling group web page:

http://bioweb.cbm.uam.es/software/MEPSA/

Version 1.0 was ported to be compatible with both Python 2.7.x and Python 3.4.x and is still available. However, to improve long term maintainability, the most recent 1.1 and 1.2 versions are only 3.4.x compatible, as any future updates will be.

3.1.3 Requirements

In general, three main packages are required to run MEPSA: The GUI was built with TKinter package[115], plotting is performed via the Matplolib package[116] and the Numpy package is extensively used for data handling and calculations[117].

MEPSA can be easily installed in Linux, Windows or Mac OS, and detailed multiplatform instructions are provided in the user manual available for download.

3.1.4 User Interface

MEPSA offers a simple GUI structured in one main window (Fig. 21 a) from which two secondary windows (i.e. "Map editor" and "Connectivity analyses") can be called (Figures 21 b and c). Additionally, the main window is used to load, unload and plot

energy surfaces, called maps in MEPSA, as well as enabling or disabling auto-plot, which determines if plots will be automatically generated after certain user actions. MEPSA supports column formatted plain text files both as input and output files which simplifies the use of custom scripts for input generation or output post-processing. Generated figures can be directly saved in many different formats (png, eps, pdf, etc.). Many GUI buttons can be found in active (black text) or inactive states (gray text). A map cannot be unloaded, plotted, edited or analyzed if it has not been loaded first. Therefore, the buttons that give access to these functions remain inactive until a map is successfully loaded. In a similar fashion there are analyses that require other ones to be done first and, consequently, are not available until such prerequisites are met.

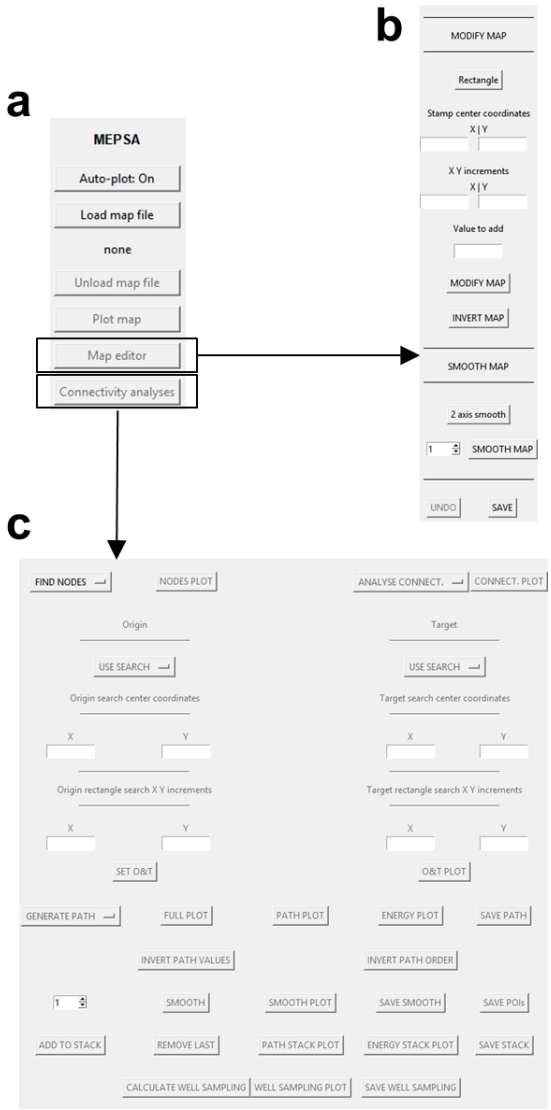


Figure 21: *MEPSA window hierarchy. From the main window (**a**) a single instance of the "Map editor" (**b**) and "Connectivity analyses" (**c**) windows can be opened as long as a map has been loaded. In the same fashion, inactive buttons in "map editor" and "connectivity analyses" widows become active as the required conditions are satisfied. For example, "ANALYSE CONNECT" and "SET O&T" buttons become active after "FIND NODES" is successfully used, while "GENERATE PATH" and "CALCULATE WELL SAMPLING" require the user to successfully select the origin and target nodes first.*

3.1.5 Functionality

The most relevant functionality MEPSA is offered by the tools grouped in "Connectivity analyses"(Fig. 21 c). These are node detection ("FIND NODES"), global connectivity analysis ("ANALYSE CONNECT."), minimum energy path generation ("GENERATE PATH") and well sampling analysis ("CALCULATE WELL SAMPLING"). There is also another small group of tools called "Map editor" which allow the different kinds of modifications over the energy surfaces.

3.1.5.1 Node detection

All MEPSA connectivity analyses consist on describing several aspects of the connectivity of nodes in an energy surface from a transition state theory perspective. Depending on the user preferences, nodes can be only minima points or minima and flat points as well. In MEPSA a flat point is defined as a point which has the same energy as its neighbors (Fig. 22). A minima point is defined as a point which has less or equal energy than all of its neighbors and less energy than at least one of them (Fig. 22).

As node detection is a requirement for all the other connectivity analyses, these remain blocked until node detection has been successfully performed.

Once nodes have been determined, they can be represented through "NODES PLOT" (Fig. 23). This plot will automatically appear after running node detection if auto-plot is enabled in the main window.

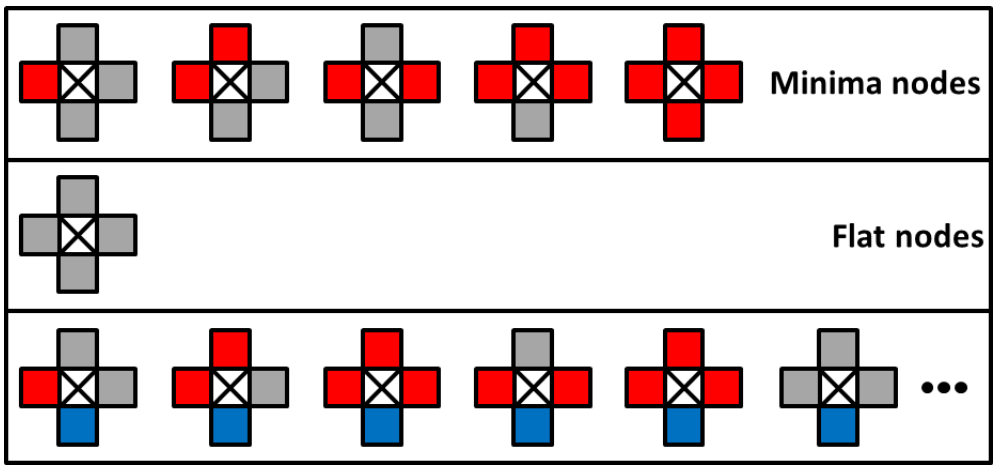


Figure 22: *Graphical depiction of node selection criteria in MEPSA. The point to be evaluated as a candidate node is indicated as a white square crossed by black lines. Red and blue squares represent points with higher and lower energy values respectively than the candidate point and gray squares represent points with the same energy. MEPSA nodes can only be minima or flat points. Therefore, any point which presents at least one adjacent point with lower energy cannot be defined as node (bottom row).*

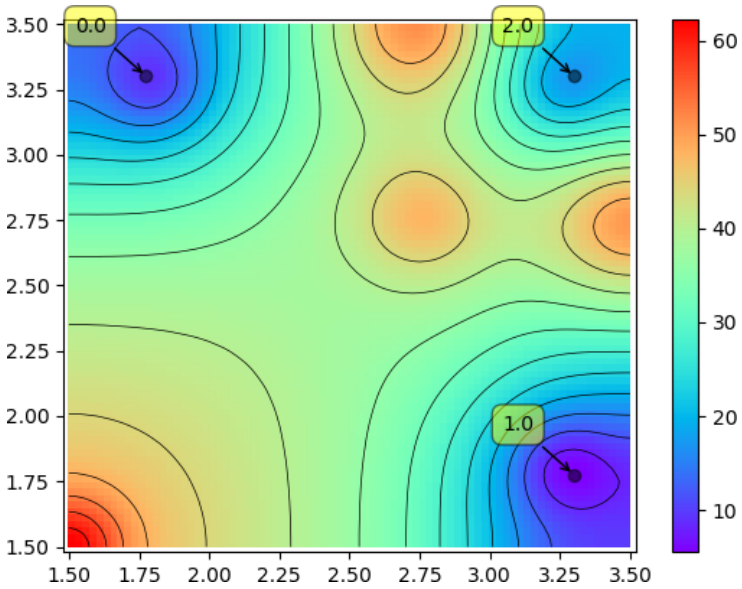


Figure 23: *Example of a node plot in MEPSA.*

3.1.5.2 Global connectivity analysis

Global connectivity analysis samples the whole map at once, starting from every detected node and iteratively propagating to the lowest energy points available until the whole surface has been sampled ("FULL" mode; Fig. 24) or every node has been connected with another one ("MINIMAL" mode; Fig. 25). On each iteration, the node id from which each propagation comes is stored to be later used to define the domain of each node. The lowest energy points located in the borders between domains (regions sloped towards a node), according to transition state theory, necessarily are the barriers connecting nodes. Therefore global connectivity analysis is able to detect the domains of each node and the energy barriers connecting them, all at once.

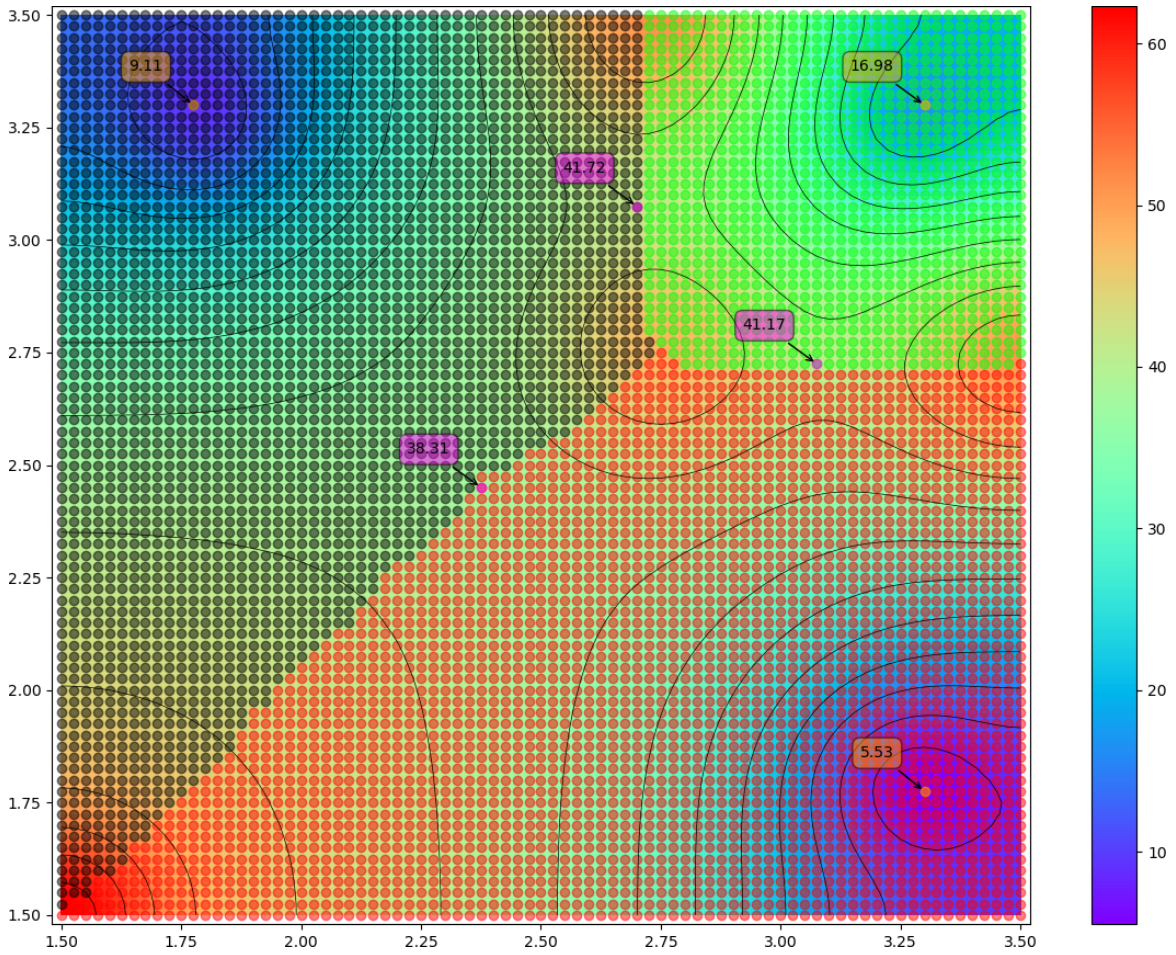


Figure 24: *Example of global connectivity analysis in "FULL" mode. The black, red and green regions represent the domain of each node, orange points indicate the nodes themselves and purple points the barriers communicating domains. Labels indicate the energy value of the annotated points.*

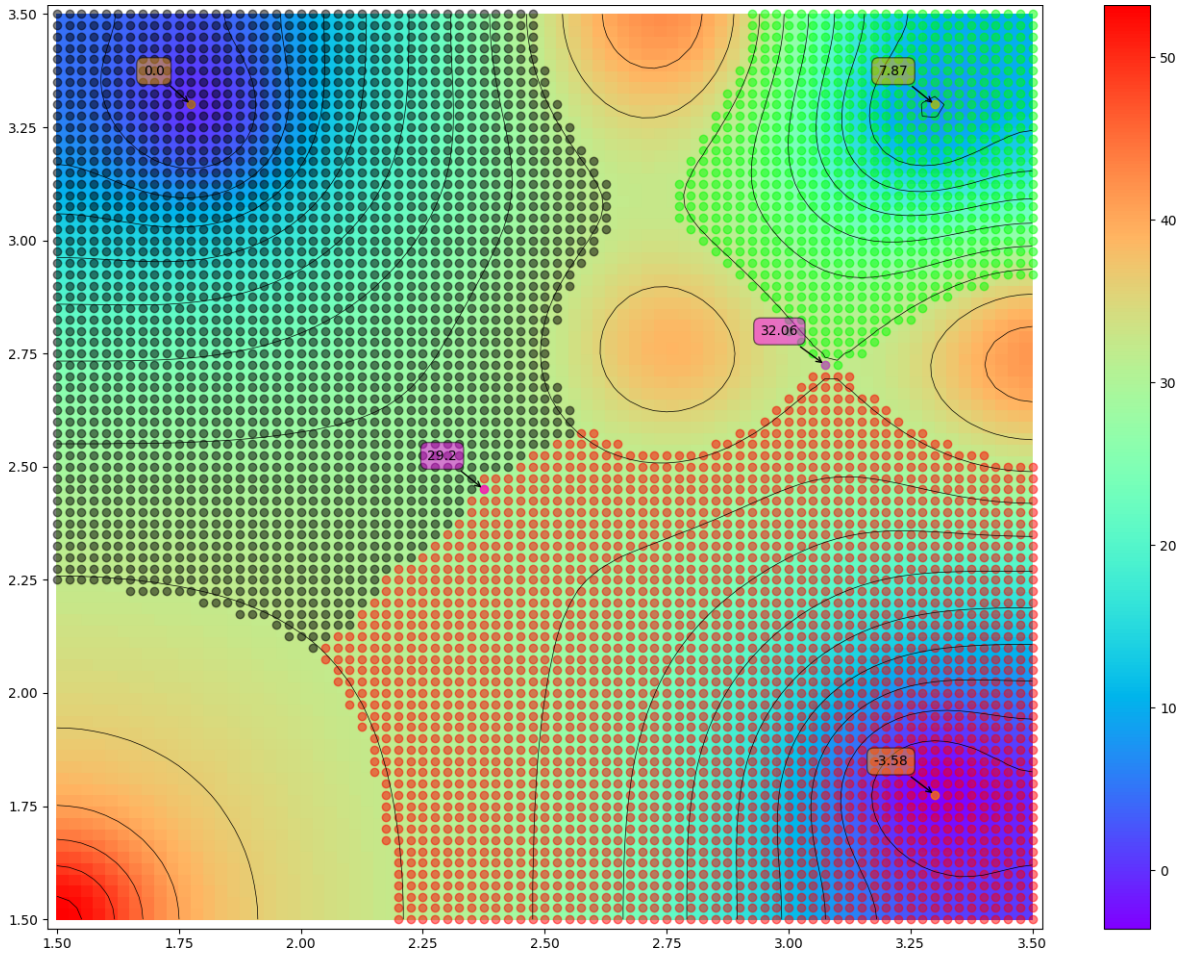


Figure 25: *Example of global connectivity analysis in "MINIMAL" mode. Symbols are as in Figure 24.*

3.1.5.3 Minimum energy path generation

After selecting an origin and target nodes (arbitrary points can be selected too but that will disable well sampling analysis) MEPSA can detect the minimum energy path between those points. The path trajectory only depends on the origin and target points selected and not on the direction, i.e. origin and target are interchangeable. There are two path generation modes: global and node-by-node. The global path-finding algorithm uses a somehow similar approach to the Dijkstra's algorithm[118] being the sampling criteria and trace back the most noticeable differences. Starting from the origin point, the program iteratively samples the surface by propagating the lowest energy points accessible until the target point is reached. On each propagation, the iteration in which the new points are occupied is stored. This information is used to perform a trace back from target to origin through the points with the lowest iteration values (Fig. 26 a). Node-by-node path-finding first runs a regular global sampling to define the order in which nodes are visited to, then, perform standard global path-finding between node pairs in that order. This way MEPSA is forced to explicitly visit every single node whose domain is visited along global path-finding (Fig. 26 b).

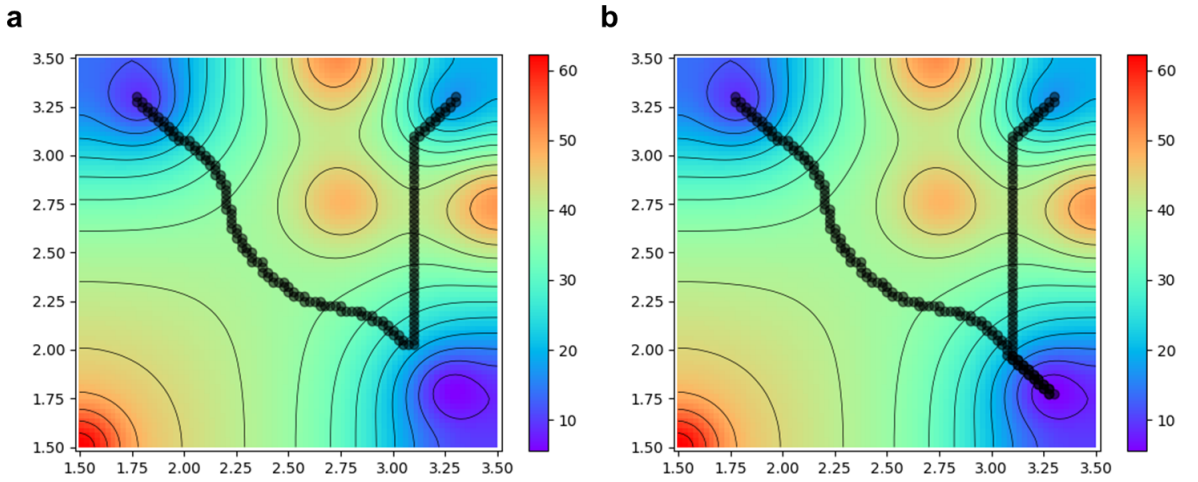


Figure 26: *Comparison of "GLOBAL" (**a**) and "NODE BY NODE" (**b**) modes of minimum energy path detection in MEPSA.*

3.1.5.4 Well sampling analysis

Once a minimum energy path has been generated, well sampling analysis becomes available. This algorithm performs a global sampling from propagates connectivity from origin and target in a similar manner to global connectivity analysis. Sampling stops upon contact of the two propagating areas, necessarily in the transition state, returning the set of points visited from origin and target to reach the transition state (Fig. 27 and 28). This analysis returns the exact region of the surface that can be meaningful for the minimum energy path detection. Among other uses this representations offers a quick view of alternative ways that were close to be sampled and, therefore, could potentially be interesting to sample in comparison (Fig. 28). As the well sampling data can be saved to a column formatted plain file, it can also be useful to generate visual representations of the relevant area of the surface (Fig. 29). Additionally, depending on the surface analyzed, this data could be potentially used to perform analyses regarding the shape and size of each well.

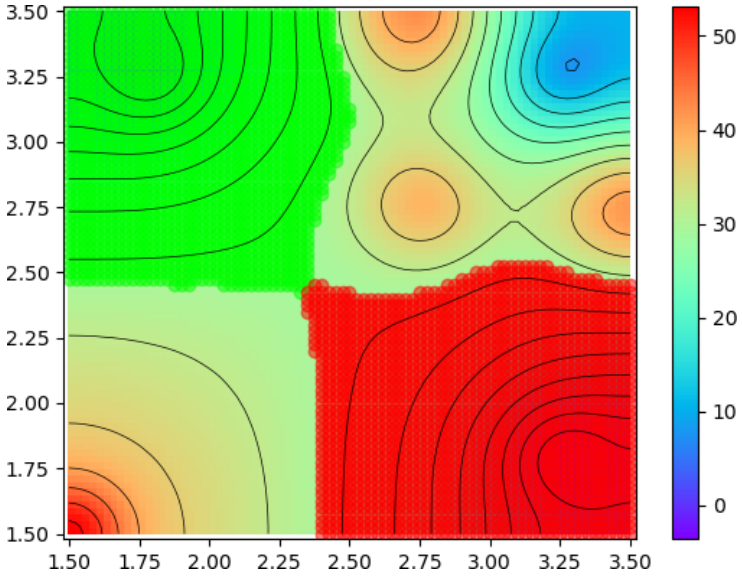


Figure 27: *Example of well sampling using nodes 0 and 1 (see Fig. 23) as origin and target.*

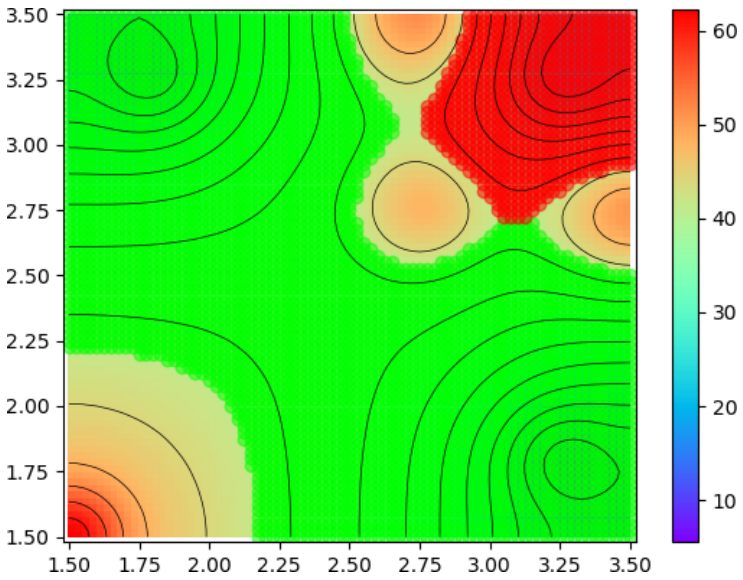


Figure 28: *Example of well sampling using nodes 0 and 2 (see Fig. 23) as origin and target. Note that barrier directly connecting nodes 0 to 2 is higher than those connecting 0 to 1 and 1 to 2. Consequently, it is not visited when well sampling is performed.*

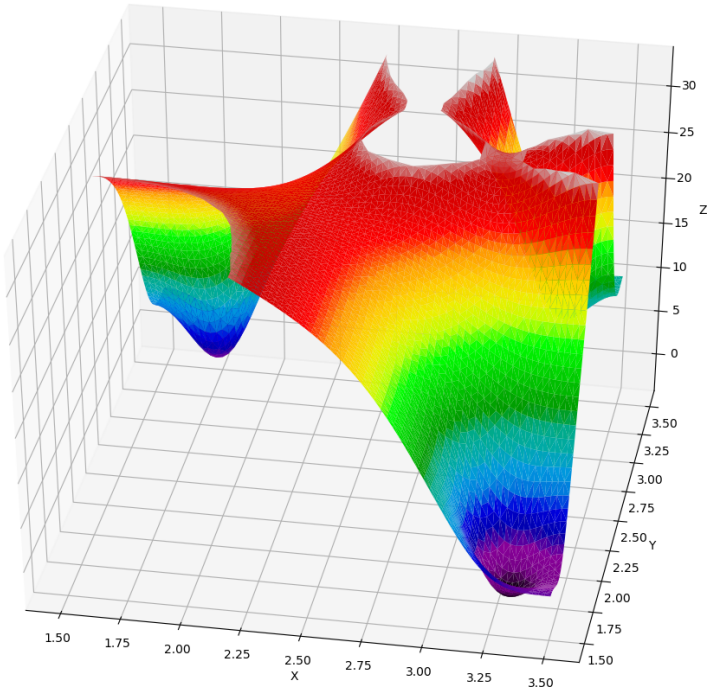


Figure 29: *3D representation by an external python script of the well sampling results presented in in figure 28. By saving MEPSA results to plain text format, further analyses and/or more flexible representations become accessible.*

3.1.5.5 Map editor

MEPSA offers some surface editing functionality grouped under "Map editor" (Fig. 21 b) which lets the user add or subtract a certain energy value over a square or circular region, allowing the generation of artificial minima and, more interestingly, the blocking of certain regions of the lowest energy path so forcing MEPSA to sample alternative paths, which can be compared with the canonical ones on the fly by using the implemented stack system in path generation. A series of paths generated under different conditions can be stored and then represented all together to visualize their differences (Fig. 30). Map editor also offers a simple running average smooth option (Fig. 31) to remove unwanted minima in noisy surfaces mainly for fast test purposes. For final results more sophisticated smooth alternatives are recommended. Lastly, a map inversion option is available, which is most useful to allow a maxima edge profile calculation in which the maximum energy values connecting to maxima are obtained by calculating a minimum energy path over an inverted surface. The results are then inverted again obtaining the profile of points with maximum energy connecting the chosen origin and target points (Fig. 32).

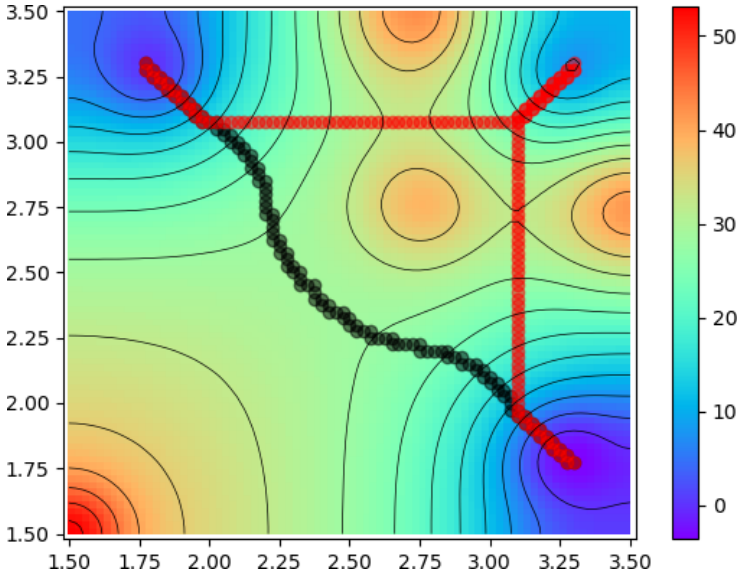


Figure 30: *Comparison of two paths. By editing the surface to force the sampling of more unfavorable barriers, alternate paths can be evaluated.*

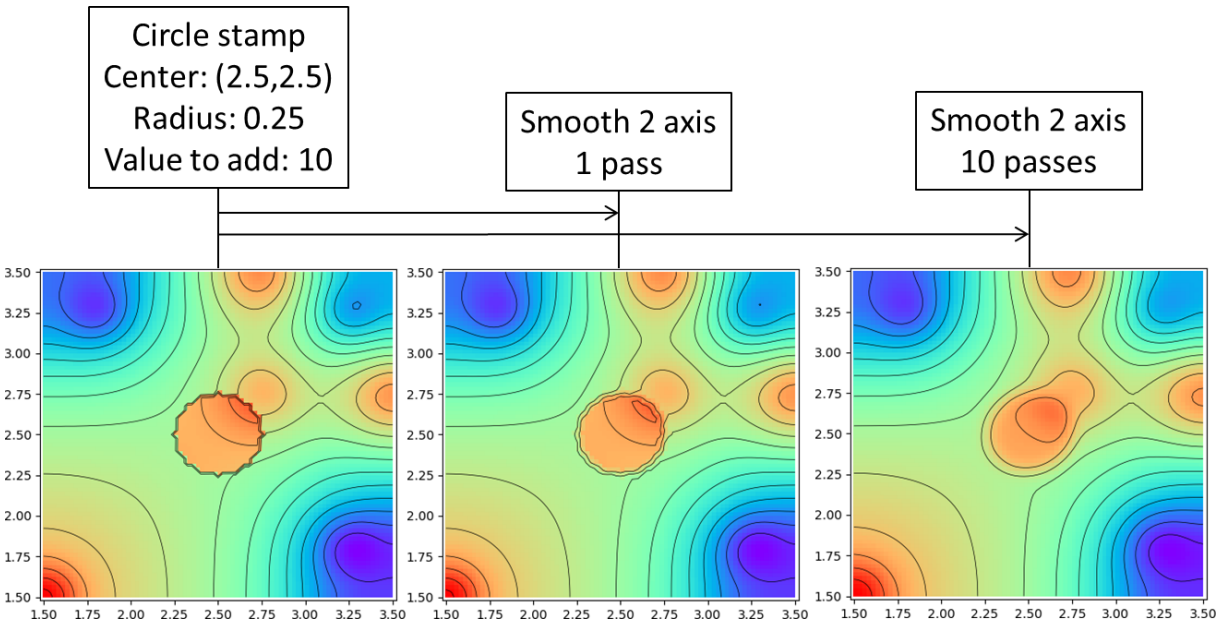


Figure 31: *Map editor smooth demonstration. A 10 arbitrary energy units circle was added to a surface which was smoothed 1 and 10 times.*

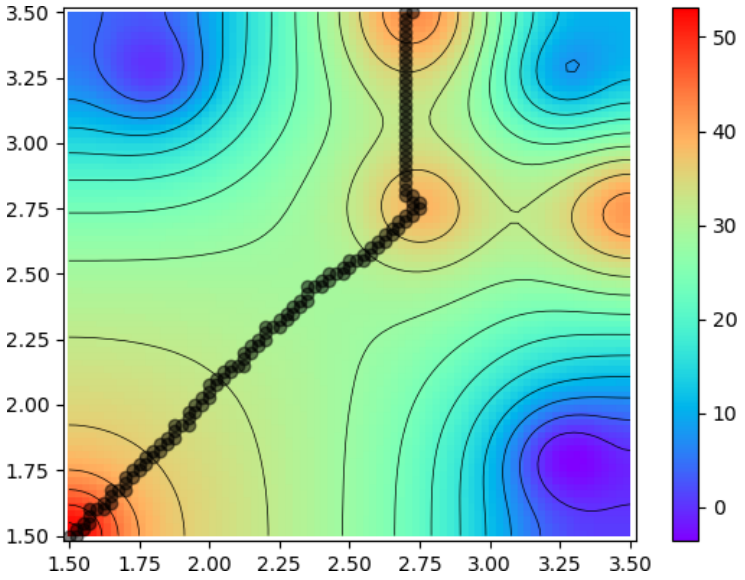


Figure 32: *Maximum edge profile. Calculating minimum energy paths over inverted surfaces allow the description of barrier profiles.*

3.1.5.6 Molecular dynamics restraints from calculated paths

The column formatted plain text files generated by MEPSA makes them very easy to parse. This can be exploited to generate a list of MD restraints that allows the user to generate an MD trajectory along a calculated path. This approach was used to generate the reaction path frames and the reaction movie describing the calculated path in figure 2, supplementary figure 1 and supplementary video 1 (appendix C) of the research paper "Two-step ATP-driven opening of cohesin head"[24], on which section 3.2 of this thesis is based.

3.1.5.7 Robustness over large and highly complex surfaces

To demonstrate that MEPSA was able to handle large and complex surfaces, a 107360 points window of topographic data covering Geneva and Turin was obtained from the United States of America NASA (National Aeronautics and Space Administration) SRTM (Shuttle Radar Topography Mission) 30 ARCSECOND ELEVATION v2.1 data set through the ORNL (Oak Ridge National Laboratory) DAAC (Distributed Active Archive Center) OGC (Open Geospatial Consortium) SDAT (Spatial Data Access Tool) webpage (https://daac.ornl.gov/spatial_data_access.shtml). The minimum elevation path between Geneva and Turin (Figures 33 and 34) and its corresponding well sampling were calculated (Fig. 35). Although some of the representations MEPSA may offer contain too much information to be useful while working with so complex surfaces, its core functionalities work and the figures they generate are still easily interpretable. As could be expected, most of the predicted minimum elevation path overlaps with drainage basins and roads (Fig. 36).

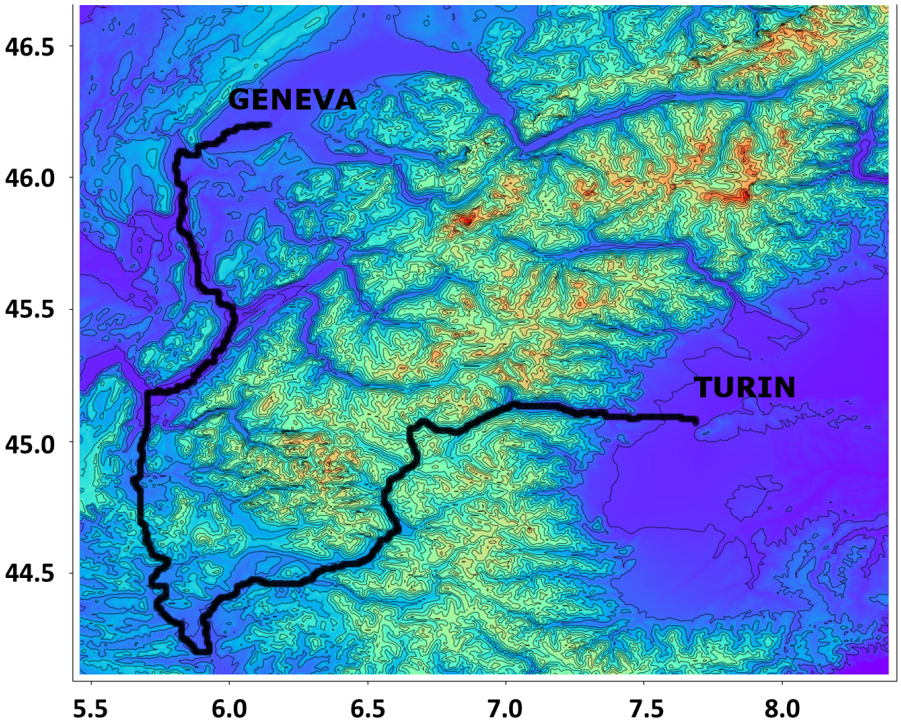


Figure 33: *Minimum elevation path between Geneva and Turin.*

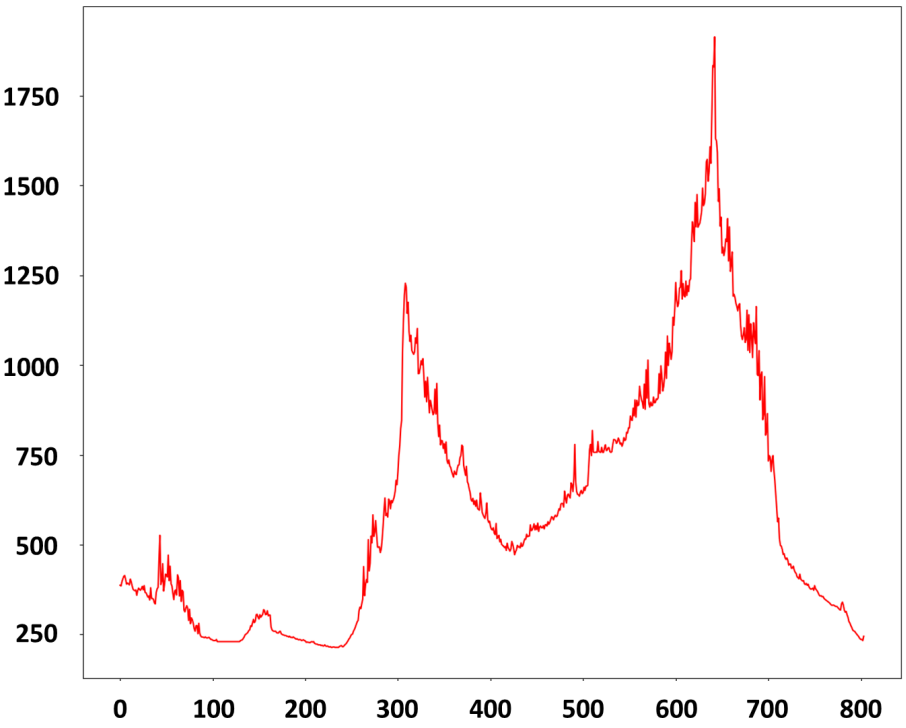


Figure 34: *Elevation profile from the path depicted in figure 33.*

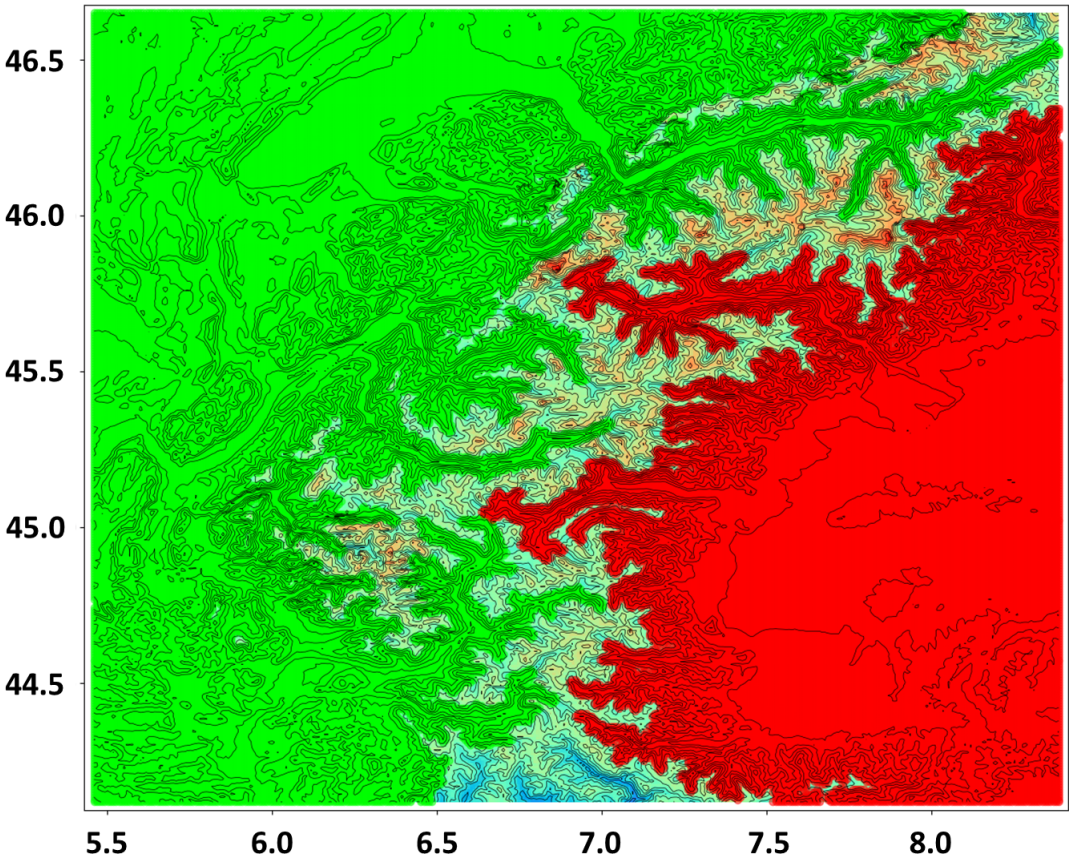


Figure 35: *Well sampling analysis of the minimum elevation path between Geneva and Turin shown in figure 33.*

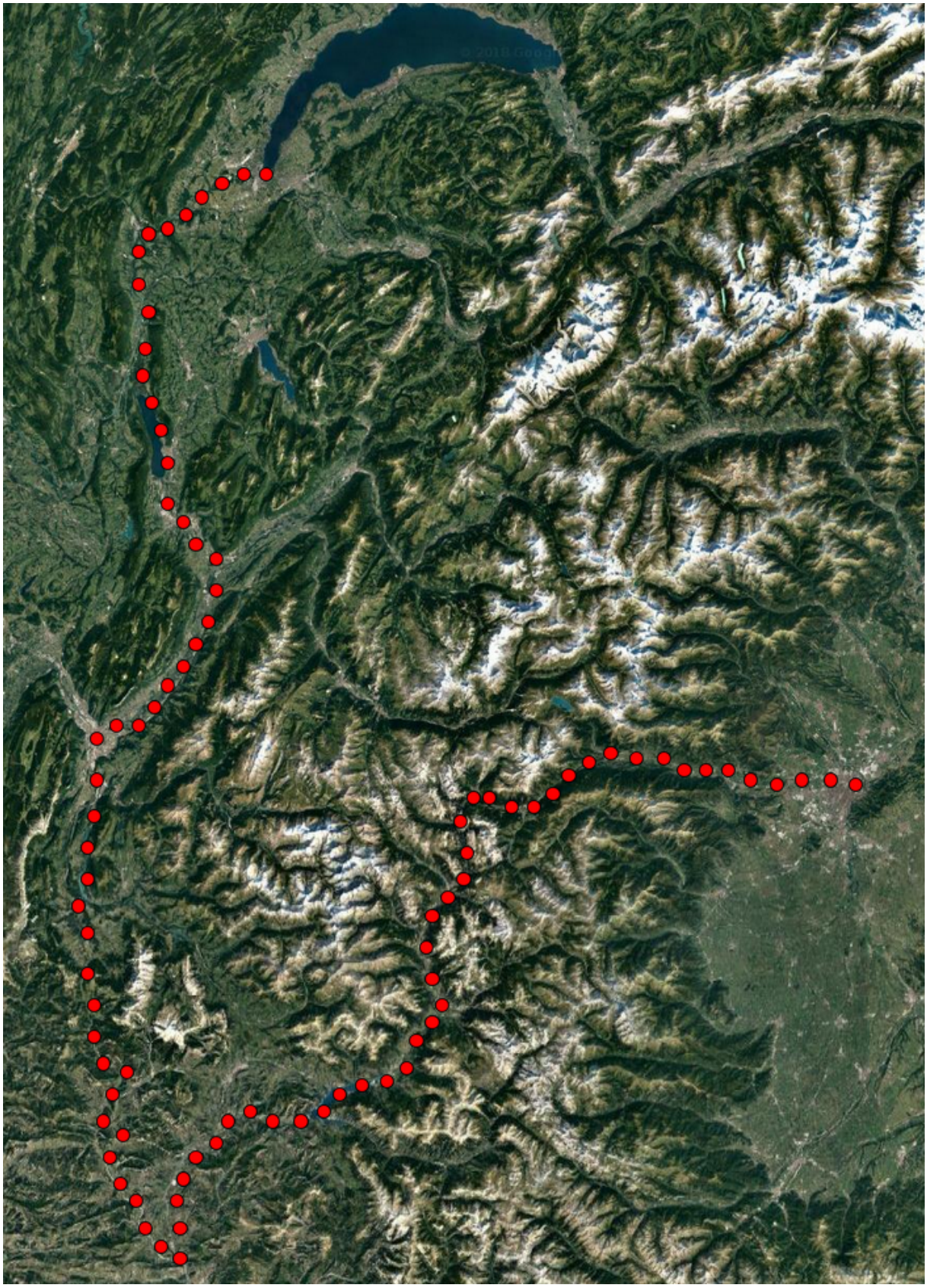


Figure 36: *Depiction of the minimum elevation path shown in figure 33 over a satellite image.*

3.1.6 Example of use

Here we will present a step-by-step MEPSA example analysis performed over a simple unitless energy map, exclusively generated for testing purposes, which can be downloaded from:

http://bioweb.cbm.uam.es/software/MEPSA/test_energy_map.zip

Multiplatform installation and run instructions are available in MEPSA manual. This example and the topographic data demonstration previously commented are explained in detailed throughout the manual. MEPSA manual can be downloaded from:

http://bioweb.cbm.uam.es/software/MEPSA/mepsa_manual_1.2.pdf

After decompression, the file "test_energy_map.dat" can be read with any plain text editor to observe the MEPSA three column input file format (if being visualized on windows, avoid using notepad as it cannot handle UNIX LF new line character). First two columns describe the two reaction coordinates and the third the energy values.

When running MEPSA the first option is to leave auto-plot enabled (default) or not (Fig. 21 a). If auto-plot is enabled, plots will be automatically generated after analyses and map editions. For the purposes of this example it is better to keep it enabled but, for heavier surfaces, it can be convenient to disable it, calling each plot only when it is actually wanted.

To load the free energy surface use the "Load map file" button (Fig. 21 a) and navigate to select the uncompressed "test_energy_map.dat" file. Once loaded, due to auto-plot being enabled, a contour plot of the surface should appear (Fig. 37). This is equivalent to press the "Plot map" button (Fig. 21 a).

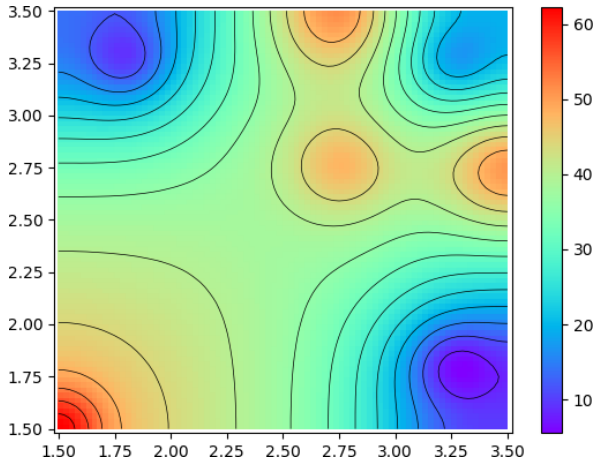


Figure 37: *MEPSA standard map plot.*

If the map has been successfully loaded, "Map editor" and "Connectivity analyses" buttons should become active.

In a typical analysis the user frequently wants to obtain the minimum energy path between two known areas of the free energy surface, e.g. the substrate and the product of a reaction, both necessarily defined by minima. Therefore, the first step to take is to search for minima. To do so, open the "Connectivity analyses" submenu, click on "FIND NODES" (Fig. 21 c) and select "MIN ONLY". A contour plot with the detected minima

annotated (Fig. 38 a) should automatically appear (this is equivalent to press "NODES PLOT" button). The associated index is the node identifier that will be used to select the origin and target nodes. In this particular example we are going to assume that node 0 corresponds to the substrate and node 1 to the product of a reaction. To calculate the minimum energy path between nodes 0 and 1, select the corresponding node identifiers from the origin and target drop-down menus and press "SET O&T" button (Fig. 21 c). Auto-plot should automatically call "O&T PLOT" if selection is successful, offering a contour plot in which origin and target nodes are depicted in green and red respectively and the energy values of both are explicitly annotated to let the user evaluate which nodes have been selected (Fig. 38 b).

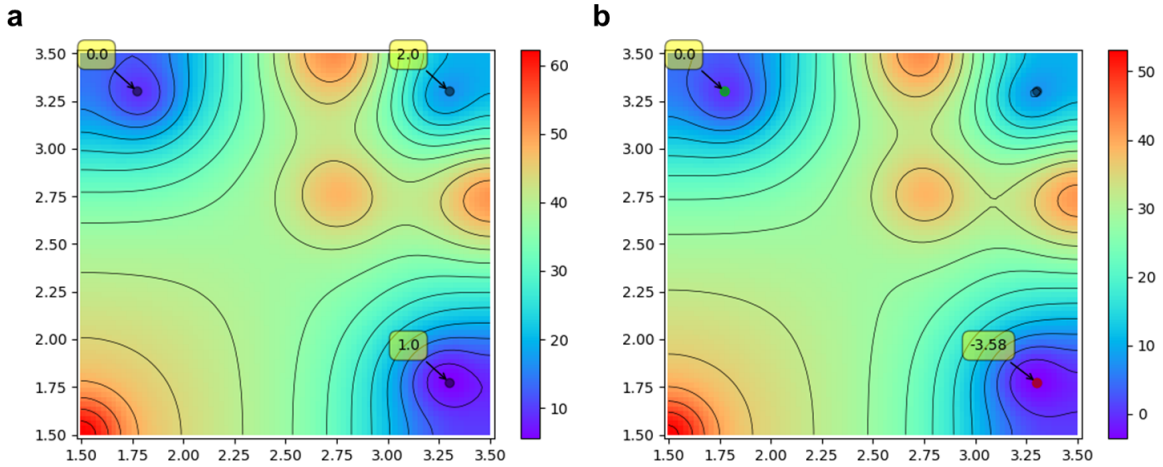


Figure 38: *MEPSA node selection. "Nodes plot" (**a**) shows the detected nodes and the indices to select them as origin or target. "O&T plot" (**b**) shows the selected origin (green) and target (red) nodes, also indicating their energy values.*

To calculate the minimum energy path between origin and target nodes click on “GENERATE PATH” (Fig. 21 c) and choose “GLOBAL”. Once complete “PATH PLOT” button is automatically called by auto-plot, showing a contour plot over which the obtained path is drawn (Fig. 39 a). “ENERGY PLOT” (Fig. 39 b) shows its energy profile and “FULL PLOT” (Fig. 39 c) offers a more complex representation in which the sampled area is highlighted in green, interest points (minima and maxima of the minimum energy path) are indicated as purple points with attached text boxes indicating energy values, minimum energy path is depicted with black points and non-visited nodes are represented as yellow points. Path can be smoothed for a number of passes specified by the user via the “SMOOTH” button (Fig. 39 d) and both the raw or smoothed paths can be saved to column formatted plain text files using their corresponding save button (“SAVE PATH” and “SAVE SMOOTH”) while “SAVE POIS” will only save the points of interest (nodes and saddle points) along the path (Fig. 21 c).

After node origin and target are successfully selected, in addition to path generation, well sampling button will become available. Well sampling can be performed by pressing the “CALCULATE WELL SAMPLING” button (Fig. 39 c), after which auto-plot should automatically call “WELL SAMPLING PLOT” (Fig.27). This plot shows a contour plot over which the points visited to reach the saddle point from target and origin nodes are highlighted. Well sampling data can be saved to a column formatted plain text file via "SAVE WELL SAMPLING" button.

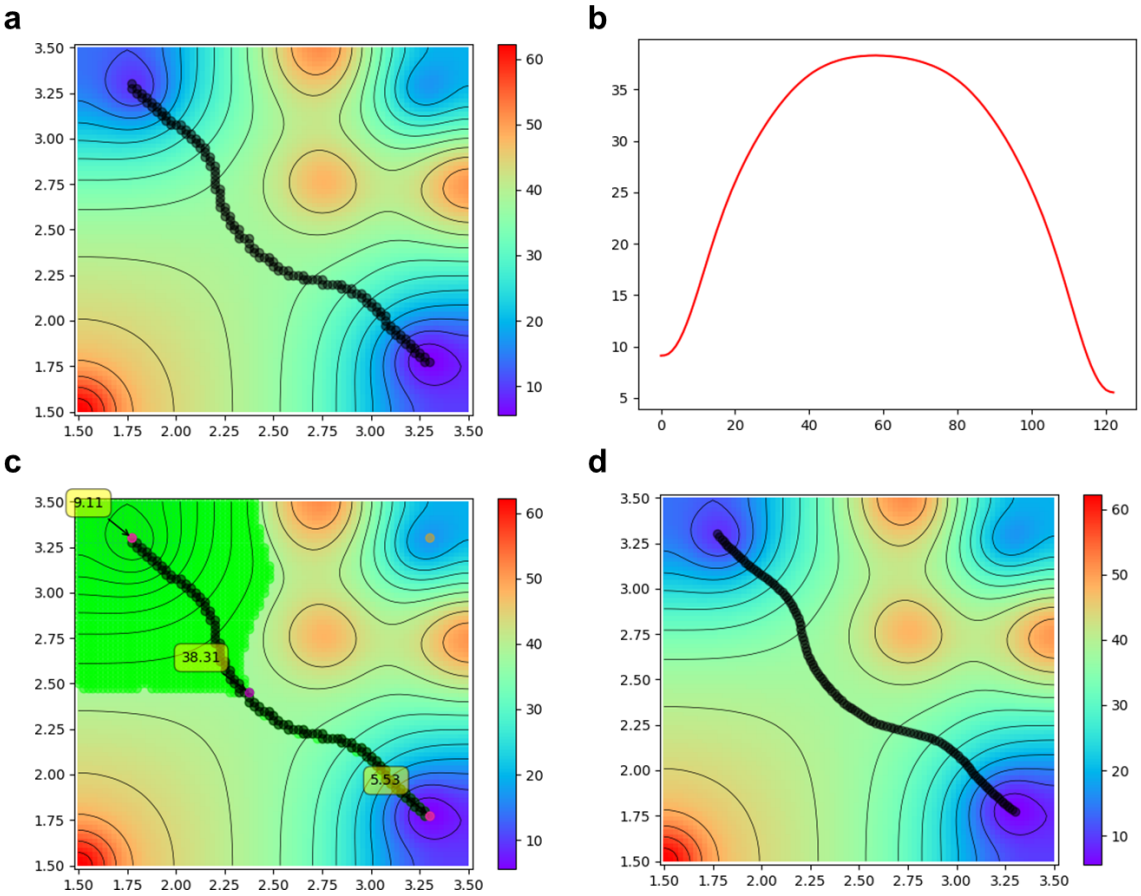


Figure 39: *MEPSA minimum energy path analysis. "Path plot" (**a**) shows the predicted minimum energy path with black dots over the energy surface. "Energy plot" (**b**) represents the energy values along the predicted minimum energy path. "Full plot" (**c**) offers a similar representation to "path plot" with additional annotation of minima and barriers location (purple dots) and energy values (yellow text bubbles) as well as the sampled region of the surface (green dots). "Smooth" option allows to average the position of adjacent points in a path the number of times specified by the user (10x in the case of figure **d**), which can later be shown via "smooth plot" (**d**).*

MEPSA also allows comparing many paths at once by using a path stack system which allows storing paths calculated under different conditions and plot them all at once. To add the calculated path to the stack press "ADD TO STACK" (21 c). Now assume we want to compare the predicted path from 0 to 1 with an alternative path passing through 2. As MEPSA will only select the minimum energy path after sampling, we need to make the current one unfavorable. To this end, map editor can be accessed from the "Map editor" button in MEPSA main window (Fig. 21 b). To disfavor the current path we need to place an artificial barrier forcing MEPSA to sample an alternative. This can be achieved, for example, by adding 10 energy units to two rectangular areas (0.5 units wide, 0.5 units long) centered in (2.5, 2.5) and (2.25, 2.25), which in the map editor is specified as a "Rectangle" stamp, with "X Y increments" X = 0.25 and Y = 0.25 (half of the desired width and length) and "Stamp center coordinates" X = 2.5, Y = 2.5 and X = 2.25, Y = 2.25 respectively, setting "Value to add" to 10. When "MODIFY MAP" button is pressed with each modification, due to auto-plot being enabled, a contour plot of the modified map should appear (Fig. 40). If the procedure previously detailed for calculating the minimum energy path is repeated selecting the equivalent nodes (which may now have different identifiers in other cases), an alternative path is obtained (Fig. 41). In order

to compare this new path with the first one, the current path has to be added to the stack by pressing the “ADD TO STACK” button (Fig. 21 c). To restore the map to its unmodified state the “UNDO” button (Fig. 21 b) in map editor window has to be used. Finally, in order to compare the two paths stored in the stack, “PATH STACK PLOT” button (Fig. 21 b) represent both paths simultaneously over a surface contour plot (Fig. 42 a) and “ENERGY STACK PLOT” button compares their associated energy profiles (Fig. 42 b).

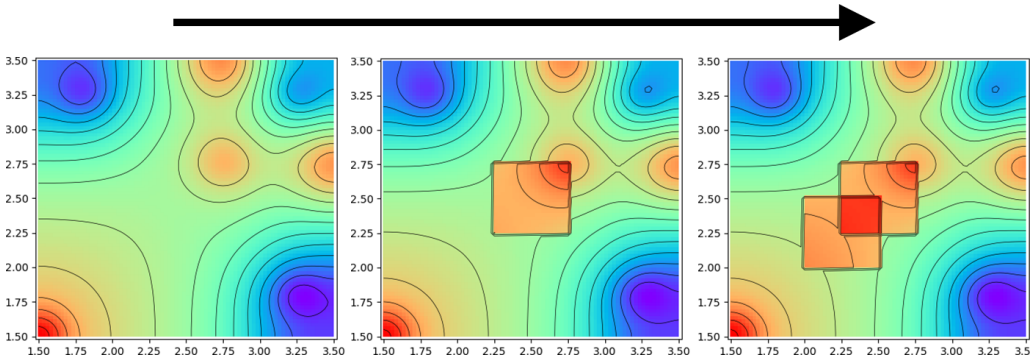


Figure 40: *Surface modification to force the sampling of alternate paths.*

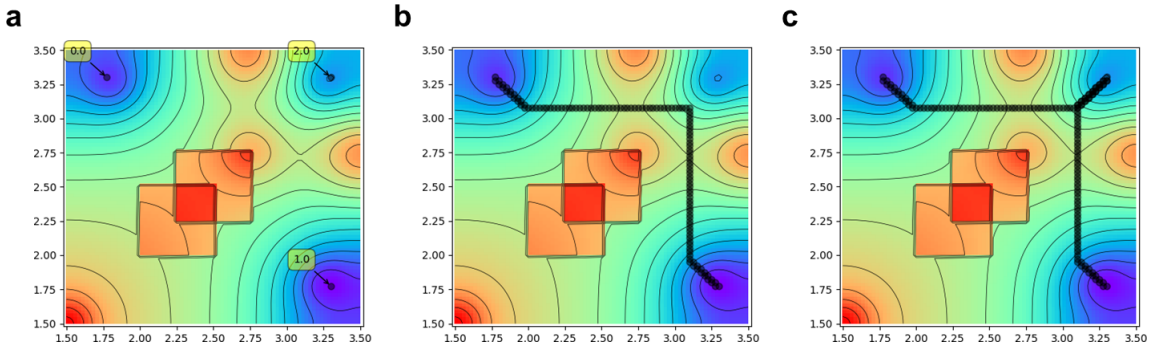


Figure 41: *Calculating an alternate path. Select nodes 0 and 1 (**a**) and calculate a global (**b**) or a node-by-node (**c**) path.*

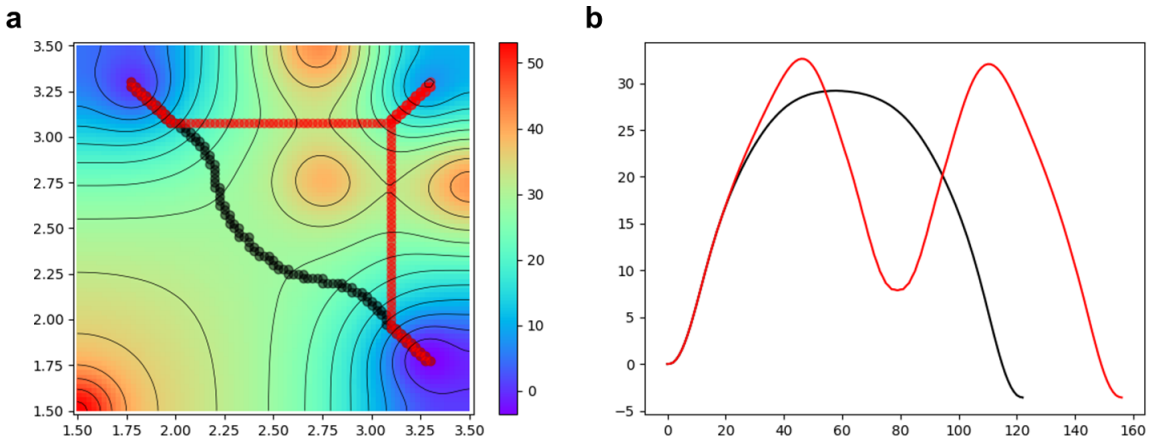


Figure 42: *Alternate paths comparison. "Path stack plot" (**a**) and "energy stack plot" (**b**) comparing the calculated paths.*

To perform a global connectivity and see all the barriers and node domains at once press “ANALYSE CONNECT.” (Fig. 21 c), selecting "FULL" or "MINIMAL" mode

afterwards. With auto-plot enabled "CONNECTIVITY PLOT" should automatically be called afterwards, showing a contour plot over which colored regions depict node domains (areas sloped towards the node), and single points represent nodes (yellow) and barriers (purple) being linked with arrows text boxes indicating their corresponding energy values (Figures 24 and 25).

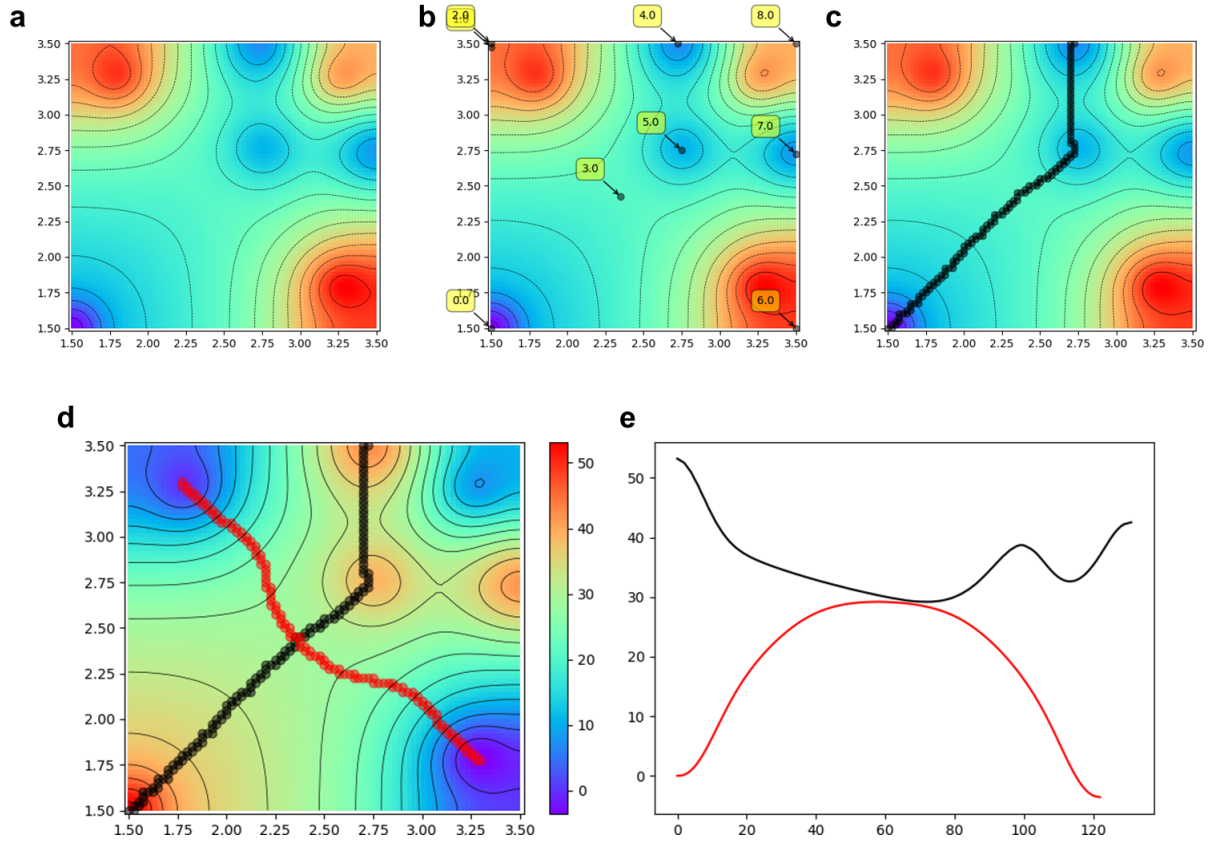


Figure 43: *Maxima edge profiling. After map inversion (**a**) and node selection (**b**) a path is calculated (**c**). The energy values of this path are inverted to restore the real values. If the maxima edge profile is to be compared with the minimum energy path, it first has to be stored in the stack. Then the minimum energy path is calculated and stored in the stack as well. Finally, the minimum energy path and maxima edge profile can be simultaneously represented via path stack plot (**d**) and energy stack plot (**e**).*

The last analysis that will be showcased is the generation of a maxima edge profile. This method is built on the principle that a maxima edge profile necessarily corresponds to a minimum energy path over an inverted surface, inverted as well after its calculation. To follow this principle, the energy surface can be inverted (Fig. 43 a) with the "INVERT MAP" button available in the map editor window (Fig. 21 b). Assuming we want to describe the maxima edge profile surrounding our substrate, we will calculate the minimum energy path between nodes 4 and 0 of the inverted surface following the method previously explained (Fig. 43 b and c). Once obtained, to invert the path back to the energy values present in the original surface "INVERT PATH VALUES" (Fig. 21 b) button has to be used. To keep the maxima edge profile in memory we will add it to the stack, but, before, we will ensure that the stack is empty. If previous paths are stored in the stack, press "REMOVE LAST" (Fig. 21) until the button becomes inactive, to ensure the stack is empty, and the press "ADD TO STACK" to store the maxima edge profile. To restore the energy surface to its unmodified state press "UNDO" in the map editor window (Fig. 21

b). Lastly, when the original surface has been brought back, use PATH STACK PLOT" and "ENERGY STACK PLOT" to visualize the calculated maxima edge profile (Fig. 43 d and e; note that in the figure the minimum energy path has been added too to offer a more illustrative perspective).

3.1.7 Discussion and perspectives

MEPSA offers a new tool to facilitate the analysis of free energy surfaces, a kind of data that is becoming increasingly accessible to researchers due to emergence of new computational methods and the increase in pure computational power. The general principles it is based on make MEPSA capable of working with any 3D surface, regardless of its complexity or the particular technique in was generated with. We adopted MEPSA as our standard method for the analysis of free energy surfaces since development started, both in its early unpublished versions[87,88] and the post-release ones[24,112], being used in every work in which we calculate free energy surfaces. Other authors have already used MEPSA[113,114] so we hope MEPSA will prove a useful tool for computational biologists, chemists and physicists in the future as the generation of 3D free energy surfaces becomes more widely used.

We keep updating MEPSA, adding some bug-fixes and improvements. However, apart from maintenance updates, long term major upgrades are being planned as we gain expertise in python development. Three are the major features planned to be added so far: complete rewrite of the core calculations to dramatically improve performance with significantly larger data sets, implement non-rectangular input data handling and, the most exciting one for us, n-dimensional generalization of MEPSA, making surfaces with an arbitrary number of dimensions analyzable with all the methods previously detailed.

3.2 Two-step ATP-driven opening of cohesin head

3.2.1 Introduction

3.2.1.1 The cohesin complex

This section will offer a general description of the cohesin complex, its relationship with disease and a list of relevant questions about it that are yet to be answered.

Function

Keeping the integrity of the genomic information over generations is crucial for any leaving being. DNA molecules containing such information in cells are highly organized through a range of macromolecular complexes with a wide variety of functions such as scaffolding, regulation of gene expression, DNA repair through homologous recombination and chromosome segregation.

To achieve faithful chromosome segregation not only the genome has to be copied with sufficient precision, but also, after successfully duplication, both resulting copies must be evenly segregated to the daughter cells.

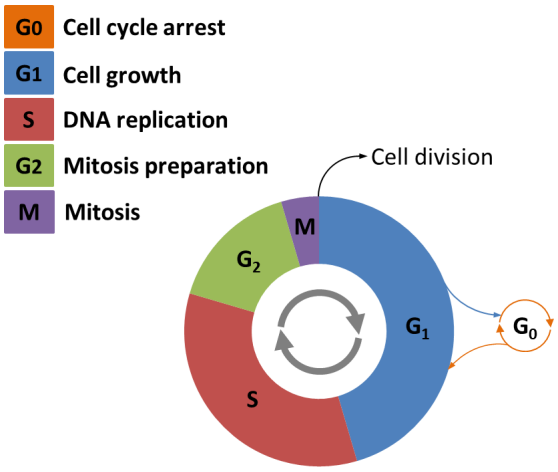


Figure 44: *Schematic depiction of the eukaryotic cell cycle.*

In the eukaryotic cell cycle (Fig. 44), in order to arrange a correct DNA distribution in the two daughter cells, chromatin has to be replicated and condensed into chromosomes, each copy of which is called a sister chromatid. Each pair of sister chromatids is held together until anaphase, when they are pulled to opposite ends of the cell prior to the actual division of the cell. Holding together sister chromatids from the beginning until they are finally ready to be moved to opposite poles of the cell provides a simple yet robust way to ensure that both copies of the same chromosome necessarily end up in a different daughter cell, requiring a minimal amount of information[119].

Cohesin ring is one of the central macromolecular complexes regulating chromosome structure. It is a protein complex capable of encircling DNA strands that plays a central role in sister chromatid cohesion (i.e. the process of holding sister chromatids together from

S phase to anaphase) and is also involved in DNA repair, chromatin organization and transcription regulation[120–123] (Fig. 45).

Chromatin loops are fundamental elements of chromosomal domain organization but the molecular mechanism through which they arise is still unclear[124]. A mechanistic model of the chromatin loops dynamics, named loop extrusion model, has recently been proposed[124–126]. Under this model, the presence of "loop extrusion factors" translocating along DNA until they interact with a "boundary element" would be sufficient to generate chromatin topologies compatible with the experimental data available[125]. It has been proposed that cohesin and condensin could be acting as "loop extrusion factors" that would translocate along the genome until they interact with the transcriptional repressor CTCF, which would be a "boundary element" under this scheme. Although this seems a rather interesting proposition, further validation is still required[124] (Fig. 45).

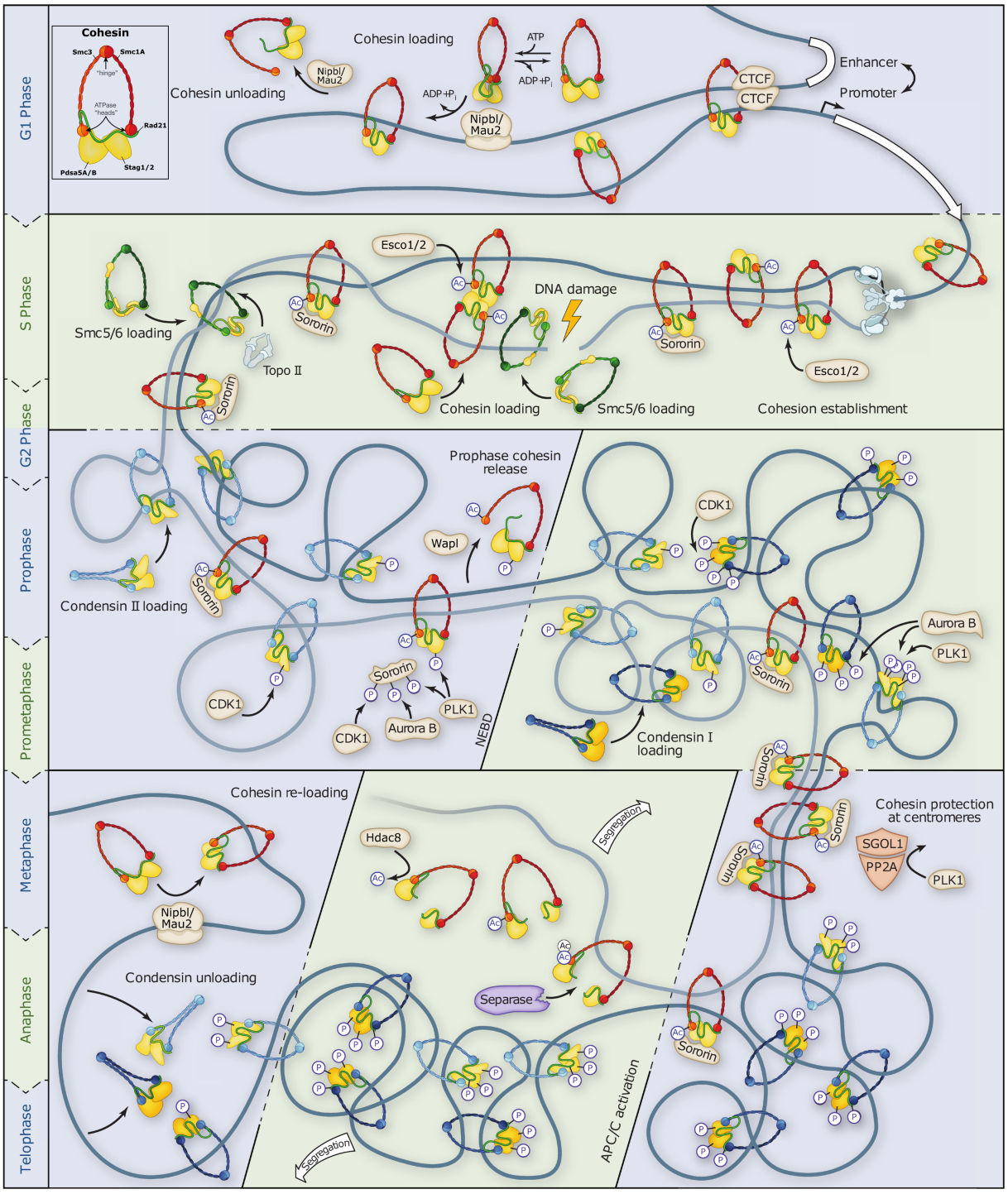


Figure 45: *Graphical depiction of cohesin complex roles along the eukaryotic cell cycle (Modified from Haering and Gruber, 2016[127]). The color code used to identify Smc1A, Smc3, Rad21, Stag1/2 and Pds5A/B is depicted in the top-left corner. The cohesin complex form rings that topologically entrap DNA. Cohesin loading is mediated by the Nipbl/Mau2 loading complex while unloading is induced by Wapl. To achieve stable cohesion during replication, a fraction of cohesin is methylated by Esco1/2, impeding Wapl-mediated unloading. Sororin is also known stabilize DNA entrapment during G_2 in several animal species. Towards the end of mitosis, cohesion is disrupted by the action of separase, which proteolytically cleaves Rad21. Putative loop extrusion expression regulation is depicted in the top-right corner, where cohesin have extruded a chromatin loop until it contacts with CTCF elements, bringing closer enhancer and promoter elements.*

Structure

The core subunits of the human cohesin complex are two structural maintenance of chromosomes (SMC) proteins, namely Smc1A and Smc3 in addition to Rad21 kleisin subunit and Stag1/2[7,120,127,128], being Smc1A, Smc3 and Rad21 the subunits that form the tripartite ring that directly encloses DNA and those on which this thesis will focus. Smc proteins are a widely conserved family of proteins from prokaryotes to eukaryotes (Fig. 2). Bacteria usually show one single Smc protein forming homodimers while eukaryotic organisms can present up to six distinct ones (in human Smc1A -or Smc1B in meiosis events-, Smc3, Smc2, Smc4, Smc5 and Smc6) forming three distinct heterodimers[129].

N-terminal and C-terminal domains of Smc proteins show conserved Walker A and Walker B motifs respectively and together create an ATPase head. As the protein folds up on itself, the structure extends along an approximately 50 nm long coiled-coil after which a dimerization domain (hinge domain) is formed.

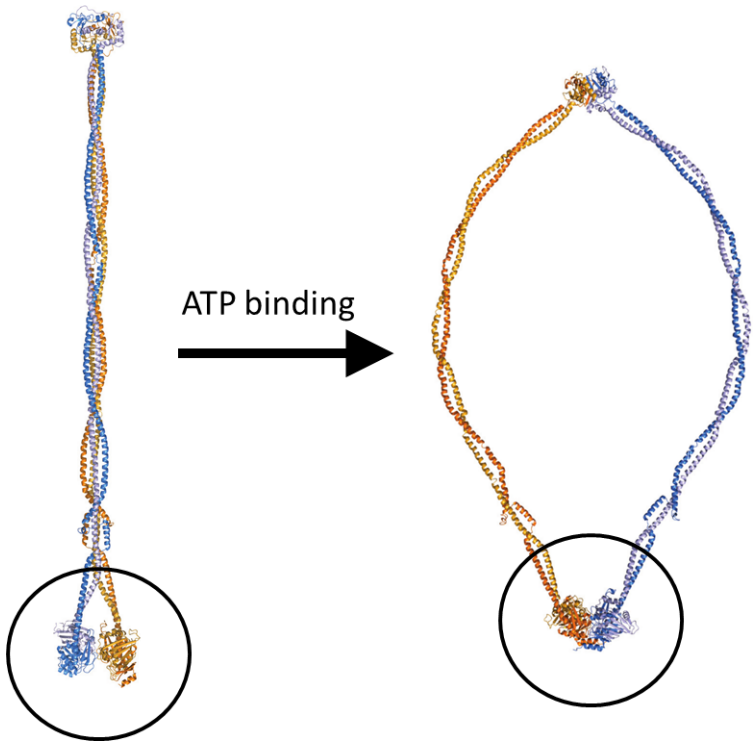


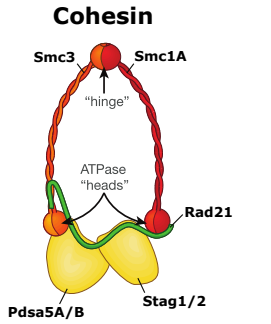
Figure 46: *Full bSMC structural model of the ATP biding dependent transition between rod-shaped and ring conformations proposed by Diebold-Durand et al. (2017)[130]. (Source: Diebold-Durand et al. (2017)[130])*

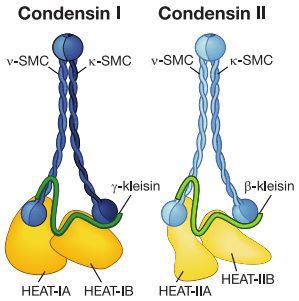
To assemble a human cohesin complex, two Smc proteins (i.e. Smc1A and Smc3) dimerize at the hinge domains and two ATP molecules get sandwiched by both ATPase heads[7], creating a ring solely closed in presence of ATP[132]. Interestingly, a structural model of the full-length prokaryotic Smc homodimer, based on data obtained from a combination of high throughput cysteine-crosslinking and crystallography, proposes a rod-like shape for bacterial cohesin in absence of ATP, being ATP binding required to induce a ring conformation[130] (Fig. 46). Although this does not have to necessarily be the case in eukaryotic cohesin, it is compatible with the known requirement of ATP binding and hydrolysis for cohesin function.

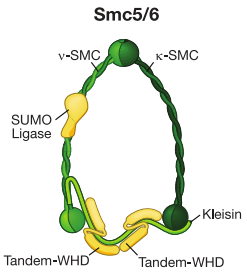
Eukaryotic SMC Complexes

		S. cerevisiae (Budding yeasts)	*S. pombe* (Fission yeasts)	*A. thaliana* (Plants)	*C. elegans* (Nematodes)	*D. melanogaster* (Insects)	*H. sapiens* (Human)
Cohesin							
Subunits	κ-SMC	Smc1	Psm1 (Smc1)	SMC1 (TTN8)	HIM-1 (SMC-1)	Smc1	**Smc1A**, SmcB*
Subunits	ν-SMC	Smc3	Psm3 (Smc3)	SMC3	SMC-3	Smc3	**Smc3**
Subunits	α-kleisin	Scc1 (Mcd1), Rec8*	Rad21, Rec8*	SYN2, SYN3, SYN4, SYN1*	SCC-1 (COH-2), COH-1, REC-8*, COH-3/4*	Rad21, C(2)M?*	**Rad21**, Rec8*; Rad21L*
Subunits	HEAT-A	Pds5	Pds5	PDS5	EVL-14	Pds5	**Pds5A**, **Pds5B**
Subunits	HEAT-B	Scc3 (Irr1)	Psc3, Rec11*	SCC3	SCC-3	SA, SA-2 (SNM)*	**Stag1/2**,Stag3*
Regulators	Kollerin (loading complex)	Scc2	Mis4	SCC2	PQN-85 (SCC-2)	Nipped-B	**Nipbl (Scc2)**
Regulators		Scc4	Ssl3	–	MAU-2	Mau-2	**Mau2 (Scc4)**
Regulators	Acetyl-transferase	Eco1 (Ctf7)	Eso1	ECO1 (CTF7)	F08F8.4	Eco (Deco) + San	**Esco1/2**
Regulators	Deacetylase	Hos1	–	–	–	–	**Hdac8**
Regulators	Stabilizer	–	–	–	–	Dalmatian	**Sororin**
Regulators	Destabilizer	Wpl1 (Rad61)	Wpl1	WAPL	WAPL-1	Wapl	**Wapl**
Regulators	Separase	Esp1	Cut1	ESP	SEP-1	Sse + Thr	ESPL1
Regulators	Shugoshin-phosphatase complex	Sgo1	Sgo1	SGO1	–	Mei-S332	SGOL1
Regulators		PP2A	PP2A	–	–	–	PP2A
Condensin							
Subunits	κ-SMC	Smc4	Cut3 (Smc4)	SMC4A	SMC-4, DPY-27**	Smc4 (Gluon)	SMC4
Subunits	ν-SMC	Smc2	Cut14 (Smc2)	SMC2A/B	MIX-1 (SMC-2)	Smc2	SMC2
Subunits	Condensin I γ-kleisin	Brn1	Cnd2	CAP-H	DPY-26	Barren (Cap-H)	CAP-H
Subunits	Condensin I HEAT-IA	Ycs4	Cnd1	CAP-D2	DPY-28	Cap-D2	CAP-D2
Subunits	Condensin I HEAT-IB	Ycg1	Cnd3	CAP-G	CAPG-1	Cap-G	CAP-G
Subunits	Condensin II β-kleisin	–	–	CAP-H2	KLE-2	Cap-H2	CAP-H2
Subunits	Condensin II HEAT-IIA	–	–	CAP-D3	HCP-6	Cap-D3	CAP-D3
Subunits	Condensin II HEAT-IIB	–	–	CAP-G2	CAPG-2	–	CAP-G2
Regulators	Cyclin-dependent kinase	Cdc28	Cdc2	–	–	–	CDK1
Regulators	Aurora B kinase	Ipl1	Ark1	–	AIR-2	Aurora B	Aurora B (AURKB)
Regulators	Polo-like kinase	Cdc5	–	–	–	–	PLK1
Smc5/6							
Subunits	κ-SMC	Smc5	Smc5 (Spr18)	SMC5	SMC-5	Smc5	SMC5
Subunits	ν-SMC	Smc6 (Rhc18)	Smc6 (Rad18)	SMC6A/B	SMC-6	Smc6 (Jnj)	SMC6
Subunits	Kleisin	Nse4 (Qri2)	Nse4 (Rad62)	NSE4A/B	–	Nse4	NSE4A, NSE4B
Subunits	Tandem-WHD E3 ligase	Nse1	Nse1	NSE1	–	Nse1	NSE1
Subunits	Tandem-WHD	Nse3	Nse3	NSE3	–	Mage (Nse3)	MAGE-G1 (NSE3)
Subunits	SUMO ligase	Mms21 (Nse2)	Nse2 (Pli2)	NSE2	–	Quijote, Cervante	NSE2
Regulators	Recruitment	Nse5	Nse5	–	–	–	–
Regulators		Kre29 (Nse6)	Nse6	–	–	–	SLF2
Regulators		Rtt107?	Brc1?	–	–	–	SLF1

*Meiosis-specific **Specific subunit of the dosage compensation complex Alternative protein names are indicated in parentheses

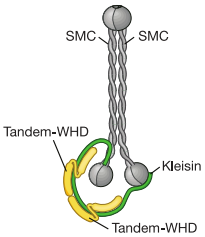






Prokaryotic SMC Complexes

		Many species (e.g. *B. subtilis*)
Smc-ScpAB		
Subunits	SMC	Smc
Subunits	Kleisin	ScpA
Subunits	Tandem-WHD	ScpB
Targeting		ParB/parS



		Sub-families of γ-proteobacteria (e.g. *E. coli*)	Var. distribution (e.g. *P. aeroginosa PAO1*)
MukBEF/MksBEF			
Subunits	SMC	MukB	MksB
Subunits	Kleisin	MukF	MksF
Subunits	Tandem-WHD	MukE	MksE
Co-factor		Topoisomerase IV	MksG

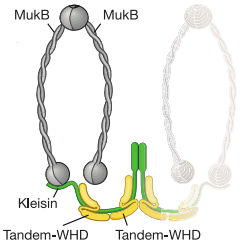


Table 2: *Conservation of the SMC protein family (Modified from Haering and Gruber, 2016[127]). SMC proteins are conserved from prokaryotes to eukaryotes and, therefore, present strong structural similarities. While in prokaryotes generally only one homodimeric complex exists, in eukaryotes three heterodimeric complexes are found: cohesin, condensin and Smc5/6. The names of the human proteins related to the cohesin complex that are discussed in this thesis (written in bold font) have been modified for clarity purposes in order to keep naming consistency.*

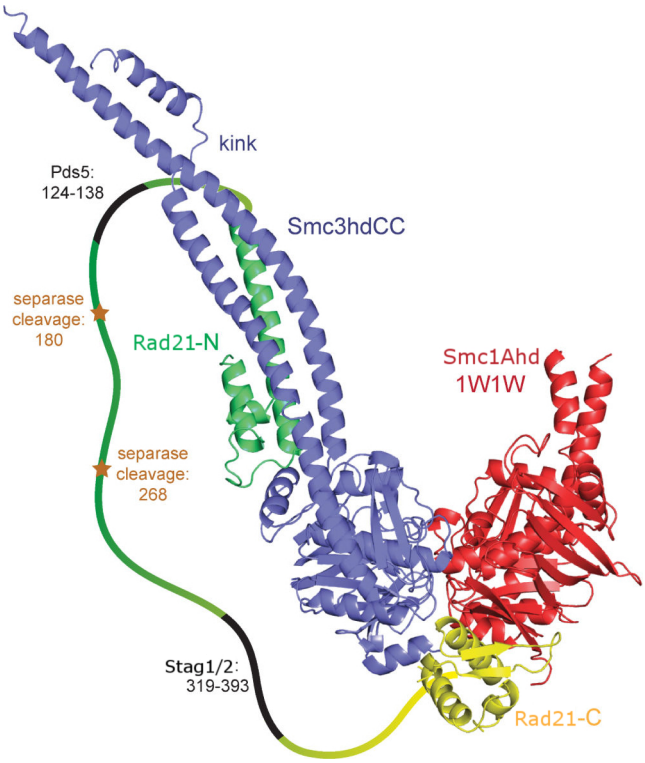


Figure 47: *Cohesin tripartite ring (Smc1A-Smc3-Rad21) model built from the 1W1W (red and yellow) and 4UX3 (blue and green) PDB structures. Protein names have been replaced by the human orthologs used in this text for clarity purposes. Modified from Gligoris et al. (2014)*[131].

Rad21 C-terminal domain binds to Smc1A ATPase head[133] while the N-terminal domain binds to Smc3 coiled-coil region adjacent to the ATPase head, forming a loose staple between both ATPase heads, thus producing a tripartite ring[131,134] (Fig. 47).

Cohesin complex constitutes a complex nanomachine powered by ATP hydrolysis, which is needed for both DNA loading[135–138] and unloading[132,139,140]. For ATP hydrolysis to effectively occur, Smc heads have to dimerize, sandwiching two ATP molecules, and interact with the C-terminal domain of a kleisin subunit (Rad21 in humans). For example, yeast cohesin Smc heads are capable of interacting in absence of Scc1 (the yeast ortholog of human Rad21) but the interaction between Smc1 (the yeast ortholog of human Smc1A) ATPase head and Scc1 C-terminal domain is required to significantly induce ATP hydrolysis[141].

Connection to disease

Mutations affecting cohesin ring elements have been related to a series of genetic disorders, known as cohesinopathies[8–13,142–144], as well as several types of cancer[14–22].

There is a variety of distinct cohesinopathies such as Roberts Syndrome, Warsaw Breakage Syndrome, CAID syndrome or CHOPS syndrome, but in this thesis we will focus on the most frequent one, Cornelia de Lange Syndrome (CdLS). CdLS is a rare developmental disorder with diverse features the severity of which may remarkably vary among affected individuals. Some of the characteristics of CdLS are: slow growth, abnormalities of the bones of arms, hands and fingers, microcephaly, intellectual disability, behavior

and neurological problems and characteristic facial features[145]. Despite there is not an exact quantification of incidence, CdLS is estimated to affect 1 in 10,000 to 1 in 30,000 newborns[145] but, due to the presence of individuals with mild or uncommon features who may never be recognized as having CdLS, the disease is believed to be underdiagnosed[146,147].

Five CdLS related genes have been described to date: NIPBL, SMC1A, SMC3, RAD21 and HDAC8[148]. Interestingly, while Hdac8 is a histone deacetylase that, in addition to the capability of deacetylation of Smc3, regulates chromatin structure and gene expression in conjunction with the rest of histone acetylases and deacetylases[149], Nipbl, Smc1A, Smc3, Rad21 all play roles in cohesin function. Smc1A, Smc3 and Rad21 form the cohesin ring[7]. Nipbl and Mau2 form a loading complex that binds chromosomes referred to as kollering[150]. Both ATP hydrolysis in the cohesin ring and interaction between the ring and kollering are required for successful topological entrapment of sister chromatids[151,152].

In addition to cohesinopathies and cancer, mutations affecting the cohesin ring have been related to aneuploidy in neurons, a relevant factor in the development of Alzheimer disease[153].

Current mechanistic models

Built upon different experimental results various mechanistic models describing certain aspects of the cohesin function coexist.

In December 2015 Yasuto Murayama and Frank Uhlmann proposed the interlocking gate mechanism model[132] (Fig. 48) that integrated the ATPase requirements for loading and unloading with a series of experimentally determined protein-protein interactions and the molecular knowledge of the acetylation regulation available at the time.

Among other strongly supported facts, they took into account that cohesin tripartite ring (Smc1A-Smc3-Rad21 complex in human) topologically entraps sister chromatids[136,154,155] and both DNA entrance and exit are mediated by ATP hydrolysis[135–138] and unloading[132,139,140]. The loading complex (Nipbl-Mau2 in human)[136,156] enhances the entrapment rate while Pds5-Wapl complex both facilitates loading and unloading of DNA[157,158]. In fission yeast the activity of both complexes depends on the presence of Psc3 (Stag1/2 in human)[136]. Interaction of DNA with the cohesin complex induces ATP hydrolysis while acetylation by Eco1 (Esco1/2 in human) of exposed lysine residues in Smc3 prevents it, leading to stable cohesion[159–164]. The schematic overview of this model can be seen in figure 48, in which the fission-yeast gene names used in the original figure by Murayama and Uhlmann have been replaced with human orthologs for clarity purposes.

In addition, in February 2016 Elbatsh et al.[139] reported four Smc1 mutations (L1129V, G1132S, D1164E and D1164G) in budding yeast that rescued viability in absence of Eco1. Complete deletion of ECO1 alleles resulted in a lethal phenotype when wild type SMC1 gene was present, while inactivation of a temperature sensitive Eco1 during S phase induced cohesion loss during metaphase. All the reported mutations were able to rescue the lethality of complete Eco1 abrogation and partial rescue of cohesion could be observed in L1129V and D1164E mutations (chosen as representative by the authors). Additionally, these mutants also exhibited a reduction of ATPase activity in presence of Scc1 (Rad21 in

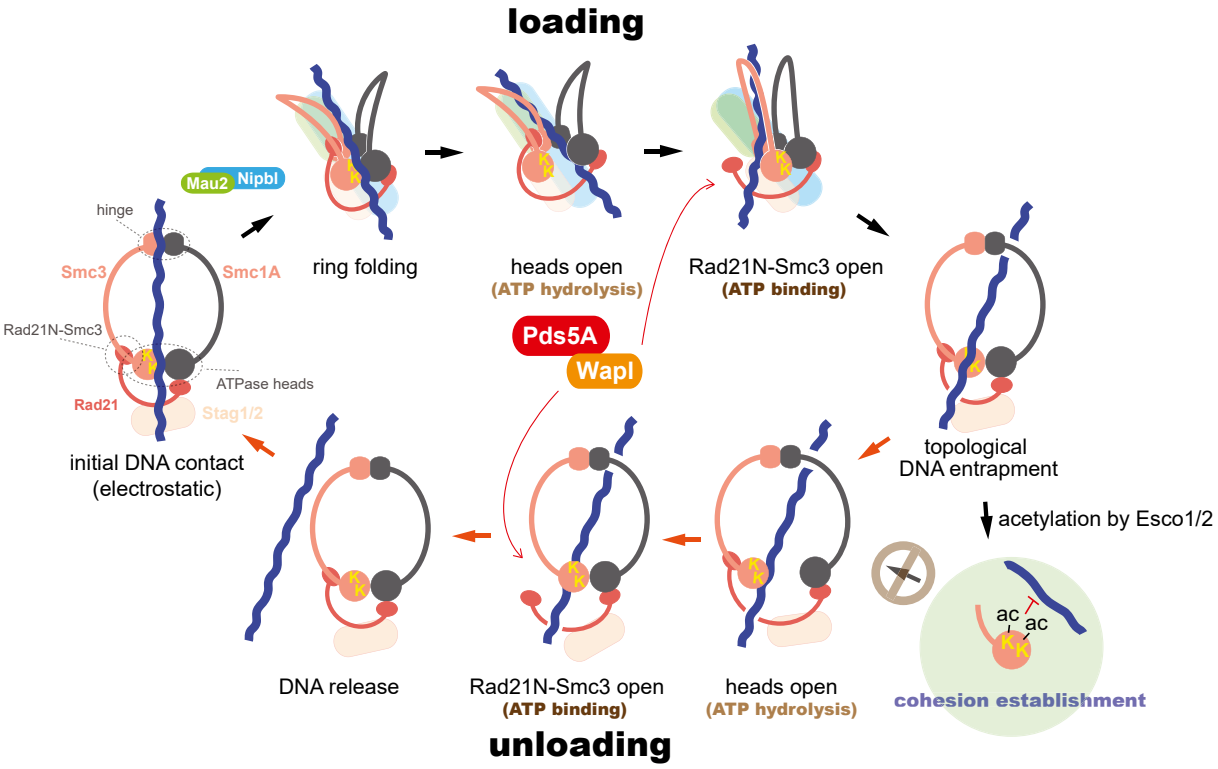


Figure 48: *Graphic depiction of the unified interlocking gate mechanism model. Protein names have been replaced by the human orthologs used in this text for clarity purposes. Modified from Murayama and Uhlmann (2015)*[132].

human) C-terminal domain similar to that induced by classical Walker B domain mutants (Smc1-E1155Q and Smc3-E1158Q). Interestingly, in contrast to the four mutants reported to bypass the need of Eco1, both Walker B mutants are lethal. Lastly, to test whether equivalent Smc3 mutation could lead to a similar phenotype, Smc3-D1161E and Smc3-L1126V mutants were evaluated, failing to reproduce the viability rescue. Among other conclusions, the authors interpret these results as a sign of a severe functional asymmetry in the core of the cohesin ATPase head.

Regarding the role cohesin could play in the loop extrusion model, in July 2017 Diebold-Durand et al.[130] proposed a model based in high-throughput crystallographic data describing an ATPase mediated DNA loop extrusion cycle sharing many similarities with the general idea presented in the interlocking gate mechanism model.

It is noteworthy that, despite these recent models try to unravel different mechanistic details of the cohesin function, they are generally compatible with each other. Moreover, they all present the ATPase dynamics undergoing in the cohesin head heterodimer as a central aspect in the cohesin function.

Open questions

Despite the attraction that the cohesin complex has caused in the scientific community over the last decades, there are still many relevant questions regarding both its structure and function are yet to be answered. Some of them deal with: the precise series of events that lead to loading, entrapment, release and stable cohesion; the exact role of the

nucleotide binding domains; and how ATP binding and hydrolysis affect the loading and releasing processes[7].

The aim of this work was to perform atomistic simulations that could offer some insight into the dynamics of the ATP hydrolysis in the active sites of the cohesin head dimer and evaluate the possible effect of such hydrolysis over the dimer stability using a model as close as possible to human cohesin so that it could possibly help to rationalize disease related mutants. The initial structure for such simulations (Fig. 49) was a homology model of the human cohesin ATPase head dimer, formed by Smc1A and Smc3 (Smc1A-head and Smc3-head respectively), bound to the C-terminal domain of human Rad21 (Rad21-Cter). The two active sites were named active site 1 (AS1) and active site 2 (AS2), being AS1 the active site formed by the Walker A and Walker B domains of Smc1A and AS2 the one formed by the Walker A and Walker B domains of Smc3 (Fig. 49). This model was simulated through MD with and without Rad21-Cter and binding different nucleotides (ATP or ADP) to each active site. The ATPase activity of both active sites was studied through QM/MM MD simulations and the free energy difference (ΔG°) between the dimer disruption before and after ATP hydrolysis was estimated via the Jarzynski's equality using a series of SMD simulations performed in presence of either ATP or ADP.

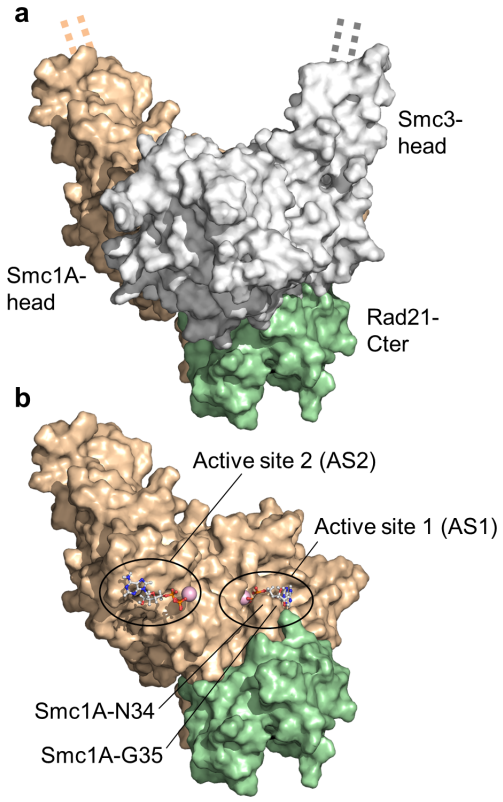


Figure 49: *Smc1A-head Smc3-head Rad21-Cter complex homology model. An overview of the homology model of the complex formed by human Smc1A-head (brown), Smc3-head (gray) and Rad21-Cter (green) is presented (**a**). To better illustrate the location of the active sites (AS1 and AS2), Smc3-head was removed (**b**). Location of the residues Smc1A-N34 and Smc1A-G35, which are in close proximity to Rad21-Cter domain, are indicated. Note their location in the active site 1 in figure 50. Source: Marcos-Alcalde et al. (2017)*[24].

3.2.2 Results

3.2.2.1 Rad21 binding induces a rearrangement at active site 1 that allows ATP hydrolysis

Rad21-Cter binding to Smc1A-head Smc3-head dimer induces ATP hydrolysis, which otherwise is almost undetectable[135,141,165].

Our hypothesis was that Rad21 is able to alter the geometry of the closest ($\sim$ 10 Å) active site (i.e. AS1) to the interface between Smc1A Rad21-Cter favoring a more catalytic configuration. In order to investigate this possible way of activation and the molecular mechanisms involved, we first performed free MD simulations of the Smc1A-head Smc3-head dimer binding ATP in both active sites, either in presence or absence of Rad21-Cter to let the model relax in each condition. To assess whether the stable conformation obtained in each condition could lead to different ATPase activity we resorted to QM/MM MD, which allows the simulation of chemical reactions such as ATP hydrolysis (see "Quantum Mechanics/Molecular Mechanics Molecular Dynamics" in subsection 1.2.3.1). To this end, the previously introduced Fireball/AMBER QM/MM MD method[87,88] was used. The computational performance of this method enables the generation of 3D free energy surfaces of enzymatic reactions without a priori determination of any reaction paths, yet providing accurate QM calculations.

After 40 ns of free MD and prior to the generation of the free energy surfaces, each condition was stabilized through 150 ps of QM/MM MD. The region described with Fireball QM calculations (QM region) (Fig. 50) was formed of the tri-phosphate moiety of the ATP, the magnesium ion, water molecules and side chains present in the coordination sphere of magnesium in AS1 as well as the catalytic water molecule and the side chains of Smc1A-N34, Smc1A-G35, Smc1A-K38, Smc1A-E1157 and Smc3-S1116. The region described with AMBER MM calculations (MM region) included the remaining atoms (i.e. the rest of the protein complex, counterions, solvent as well as AS2 ATP and magnesium ion). After QM/MM MD stabilization free energy surfaces were sampled for both conditions along two reaction coordinates, reaction coordinate 1 (RC1) and reaction coordinate 2 (RC2; purple arrows in figure 50). RC1 was defined to describe the bond to be formed, i.e. the distance between the oxygen atom from the catalytic water molecule and the phosphorus atom of the ATP γ-phosphate group. RC2 was defined to describe the bond to be broken, i.e. the internal distance in the ATP molecule between the phosphorus atom of the γ-phosphate group and the third oxygen atom of the β-phosphate group. For each condition, this sampling yielded 7.6×10^6 conformations with their corresponding reaction coordinates and QM energy values, all of which were used to generate the free energy surfaces present in figures 51 and 52. The minimum free energy paths (cyan line in figure 51 a and red line in figure 52 a for each surface were calculated using MEPSA[23]. The energy profiles of those paths (figures 51 b and 52 b) show the free energy evolution along the path from the initial ATP molecule (figures 51 c "1(S)" and 52 c "1(S)") to the resulting ADP and inorganic phosphate molecules (figures 51 c "6(P)" and 52 c "6(P)"). Interestingly both conditions exhibit rather similar paths, maxima and minima locations and almost equivalent ΔG° differences between substrate and product but, on the other hand, large differences in the free energy of activation ($\Delta^\ddagger G^\circ$) can be observed (Fig. 53).

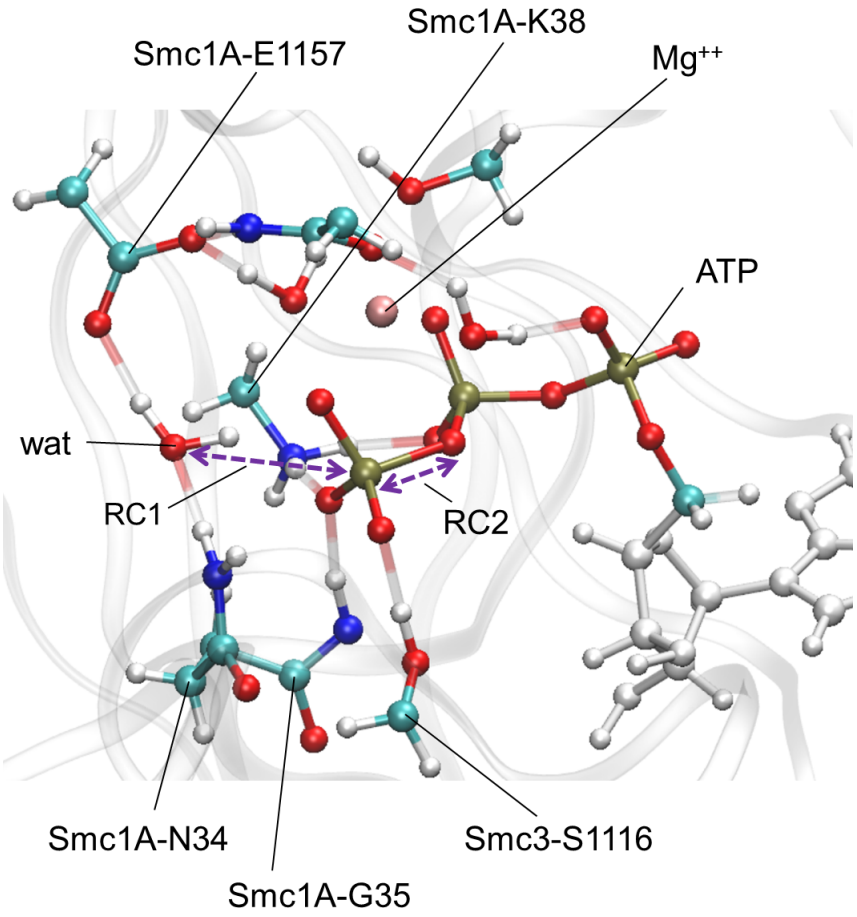


Figure 50: *QM region for AS1. Atoms belonging to the QM region are depicted with colored ball and sticks. Atoms represented in gray belong to the MM region. Light-gray ball and sticks conform the rest of the ATP molecule that is not contained in the QM region while the background transparent gray ribbons represent the a-carbon trace of the protein chains surrounding the QM region. The positions of the catalytic water (wat), residues Smc1A-N34, Smc1A-G35, Smc1A-K38, Smc1A-E1157 and Smc3-S1116, magnesium ion (Mg^{++}) and ATP molecule are indicated. Reaction coordinates 1 (RC1) and 2 (RC2) are symbolized by purple arrows. Source: Marcos-Alcalde et al. (2017)[24]. Caption was adapted from the same source.*

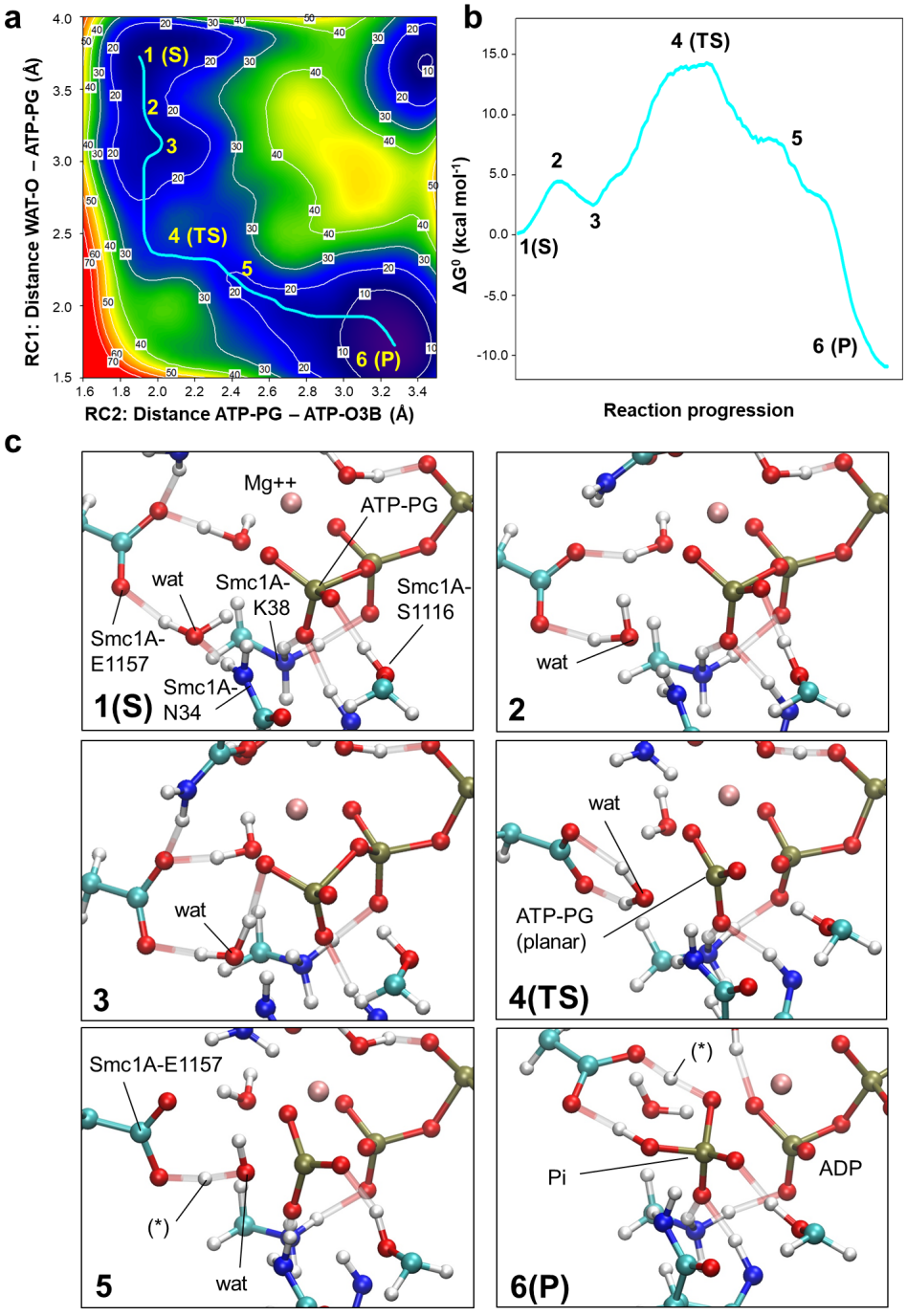


Figure 51: *Free energy surface analysis of ATP hydrolysis in AS1 in presence of Rad21. (**a**) Free energy surfaces (in kcal mol^{-1}) for ATP hydrolysis at AS1 in the presence of Rad21-Cter generated via QM/MM MD simulations. The plot axes represent the reaction coordinates. RC1 (bond to be formed): the distance (in Å) between the oxygen atom of the catalytic water and the phosphorous atom of the ATP molecule γ-phosphate group (distance wat-O- ATP-PG). RC2 (bond to be broken): the distance (in Å) between the phosphorous atom of the ATP molecule γ-phosphate group and the oxygen atom 3 of the ATP β-phosphate group (distance ATP-PG - ATP-O3B). Free energy data are represented via a color scale, from lower (blue) to higher (red) values. MEPSA minimum energy path is shown in cyan. (**b**) Free energy profile of the MEPSA minimum energy path. Points of interest (1(S), 2, 3, 4(TS), 5, 6(P)) are indicated. (c) Representative structures of the points of interest revealed by the free energy profile are shown. The positions of the catalytic water (wat), residues Smc1A-N34 and Smc1A-E1157, magnesium ion (Mg^{++}), ATP γ-phosphate (ATP-PG), ADP and leaving inorganic phosphate (Pi) are indicated. Source: Marcos-Alcalde et al. (2017)[24]. Caption was adapted from the same source.*

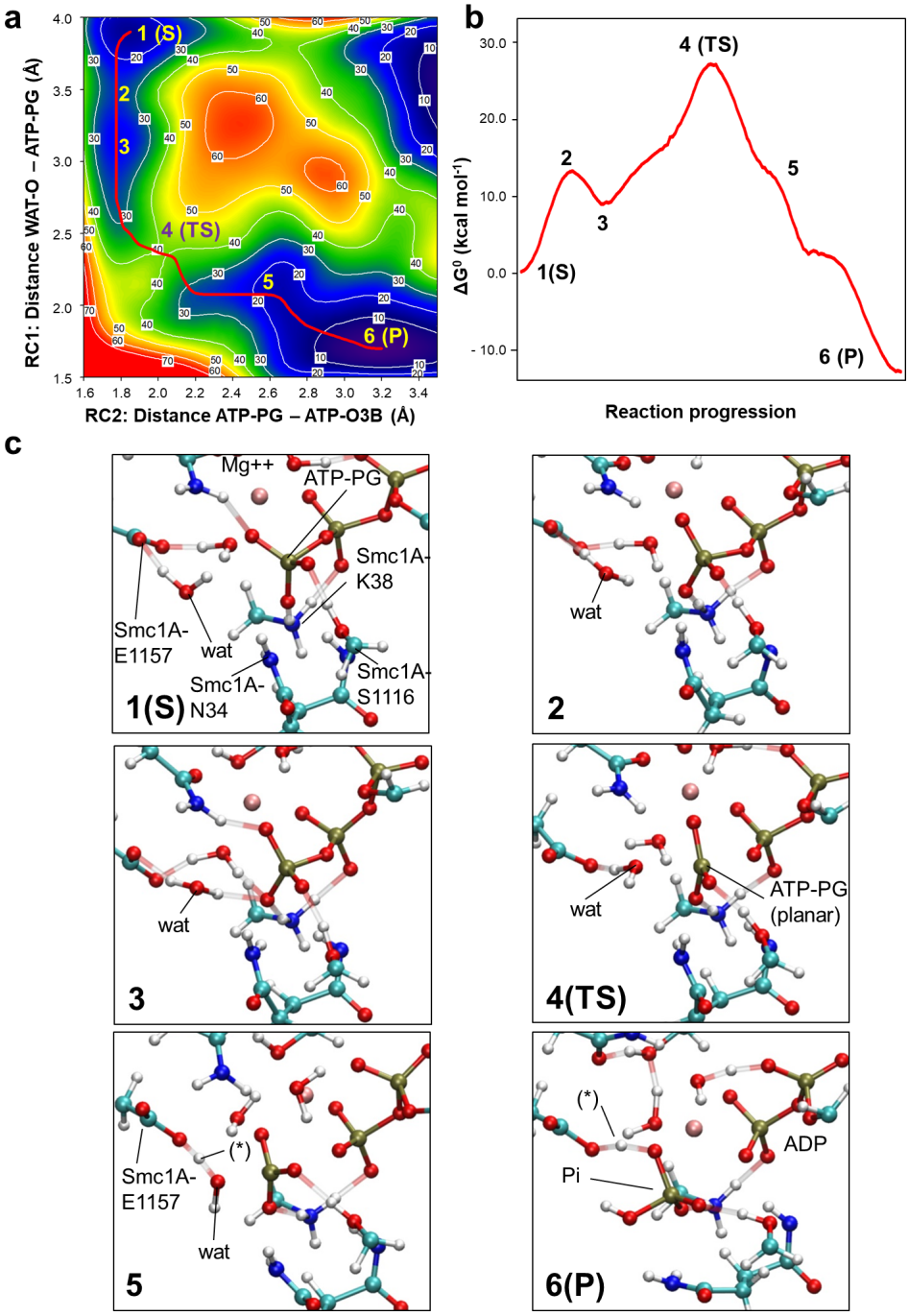


Figure 52: *Free energy surface analysis of ATP hydrolysis in AS1 in absence of Rad21. (**a**) Free energy surfaces (in kcal mol^{-1}) for ATP hydrolysis at AS1 in absence of Rad21-Cter generated via QM/MM MD simulations. The plot axes and color scale are as in figure 51 a. MEPSA minimum energy path is shown in red. (**b**) Free energy profile of the MEPSA minimum energy path. Points of interest (1(S), 2, 3, 4(TS), 5, 6(P)) are indicated. (c) Representative structures of the points of interest revealed by the free energy profile are shown. The positions of the catalytic water (wat), residues Smc1A-N34 and Smc1A-E1157, magnesium ion (Mg^{++}), ATP γ-phosphate (ATP-PG), ADP and leaving inorganic phosphate (Pi) are indicated. Source: Marcos-Alcalde et al. (2017)*[24]*. Caption was adapted from the same source.*

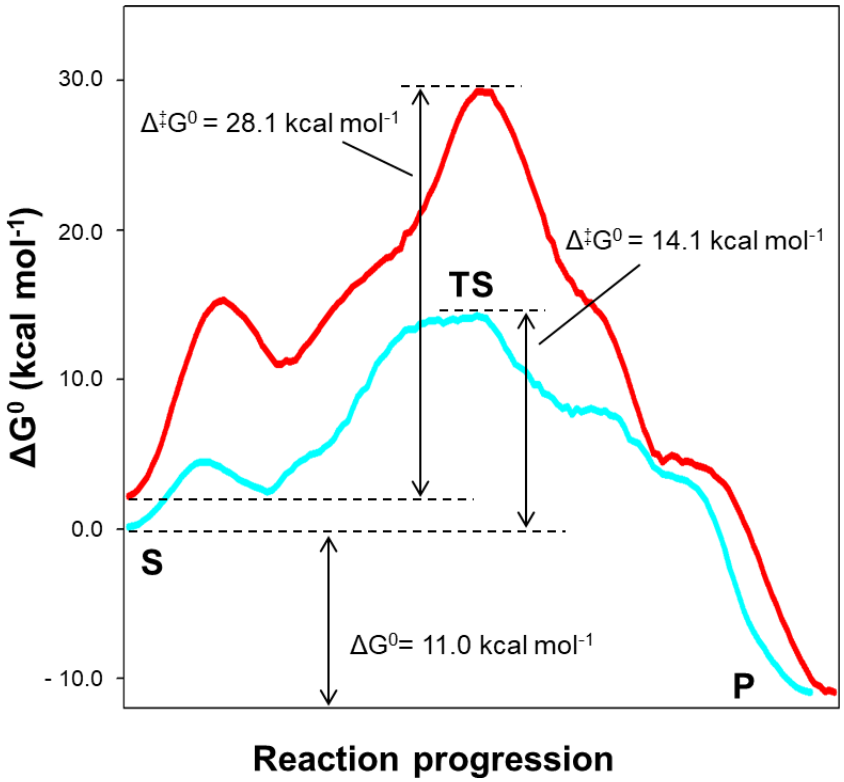


Figure 53: *The binding of Rad21-Cter to Smc1A-head facilitates ATP hydrolysis. A comparison of the free energy profiles of ATP hydrolysis in AS1 in presence (cyan) and absence (red) of Rad21-Cter is shown, indicating the location of the substrate (S), transition state (TS) and product (P). Activation free energy ($\Delta^{\ddagger}G^{\circ}$) for both condtions and free energy difference between susbtrate and product(ΔG°) in presence of Rad21-Cter are shown. Source: Marcos-Alcalde et al. (2017)*[24].

This suggests that, either in presence or absence of Rad21-Cter, the reaction takes place through the same series of steps but, when Rad21-Cter is present, some of them are significantly stabilized. To identify the possible reason for such stabilization, representative structures of each maximum and minima were compared between conditions. The two first maxima indicate the two major steps of the reaction, water entrance (figures 51 c "2","3" and 52 "2","3") and pyrophosphate bond breaking through a planar transition state (figures 51 c "4(TS)" and 52 c "4(TS)"). The comparison of structures from both conditions revealed that, in presence of Rad21-Cter, Smc1A-N34 strongly stabilizes water entrance and, to a lesser extent, the transition state, which is mainly stabilized by Smc1A-K38 and Smc1A-G35. Interestingly, Rad21-K605 directly contacts with Smc1A-G35 and maintains Smc1A-N34 in position by interacting with Smc1A-G35, offering a possible way to rationalize the molecular mechanism through which Rad21-Cter binding may alter the ATPase activity in AS1. To complete the structural description of the features associated with the energy profile, in presence of Rad21-Cter, after the transition state, a subtle energy shoulder can be observed. Representative structures of this region suggest that it is associated with the water deprotonation event prior to the bond formation (figures 51 c "5" and 52 c "5"). A video sequence of the reaction along the minimum energy path in presence or Rad21-Cter, highlighting water stabilization, transition state and water deprotonation, is shown in appendix C.

From an energetic perspective, the free energy of activation detected for ATP hydrolysis

in AS1 was 14.1 kcal mol^{-1} in presence and 28.1 kcal mol^{-1} in absence of Rad21-Cter, thus implying a 14.0 kcal mol^{-1} difference between both barriers. These results are in agreement with the experimentally observed fact that the presence of Rad21-Cter allows ATP hydrolysis[135,141,165] lowering the barrier to a value close to the range of the experimental free energy barrier measured for other ATPases, as the F1-ATPase (12.9–13.4 kcal mol^{-1})[166].

3.2.2.2 ATP hydrolysis at active site 1 induces the activation of site 2

After simulating ATP hydrolysis in AS1 we tested if this could have any detectable effect over Smc1A-head Smc3-head dimer. To such end, the Smc1A-head Smc3-head Rad21-Cter model binding ADP in AS1 and ATP in AS2 (AS1-ADP/AS2-ATP condition) was subjected to 150 ns of free MD. As a control, the same model, binding ATP in AS1 and ATP in AS2 (AS1-ATP/AS2-ATP condition), was subjected to an equivalent simulation. The movement of residues in the moieties of both active sites was monitored throughout both trajectories. Surprisingly, after 120 ns of free MD the AS1-ADP/AS2-ATP showed a noteworthy behavior which could not be observed in AS1-ATP/AS2-ATP. The side chain of an apparently non-related residue, Smc1A-K1120, moved into AS2 and remained stable in such position. Nitrogen of the α-amino group of Smc1A-K1120 in AS1-ADP/AS2-ATP remained stable at around 2.5 Å of the oxygen atom of the water molecule (Fig. 54) while, in AS1-ATP/AS2-ATP, this distance fluctuated around 7.9 Å (Fig. 54).

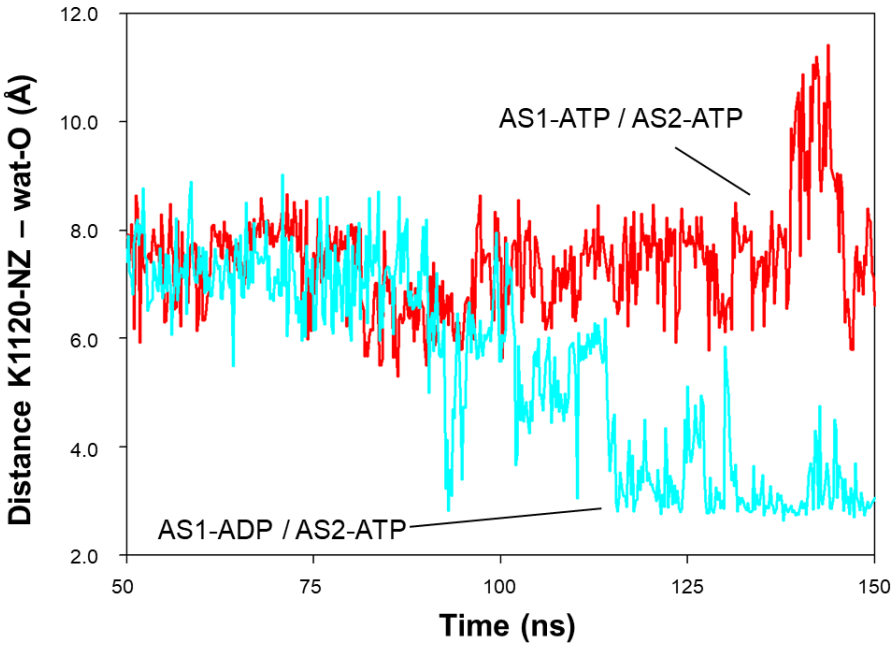


Figure 54: *After ATP hydrolysis in AS1, Smc1A-K1120 approximate and contacts the catalytic water molecule (wat) in AS2. The evolution of the distance between the oxygen of wat and the ϵ-amino group of the Smc1A-K1120 residue (distance K1120-NZ - wat-O) is shown for two trajectories, one in which both active sites were binding ATP (AS1-ATP/AS2-ATP; red line) and one in which AS1 was binding ADP and AS2 ATP (AS1-ADP/AS2-ATP; cyan). Source: Marcos-Alcalde et al. (2017)*[24].

In this configuration, after the entrance of Smc1A-K1120 in AS2 under AS1-ADP/AS2-ATP condition, the α-amino group of the lysine formed stable hydrogen bonds with the

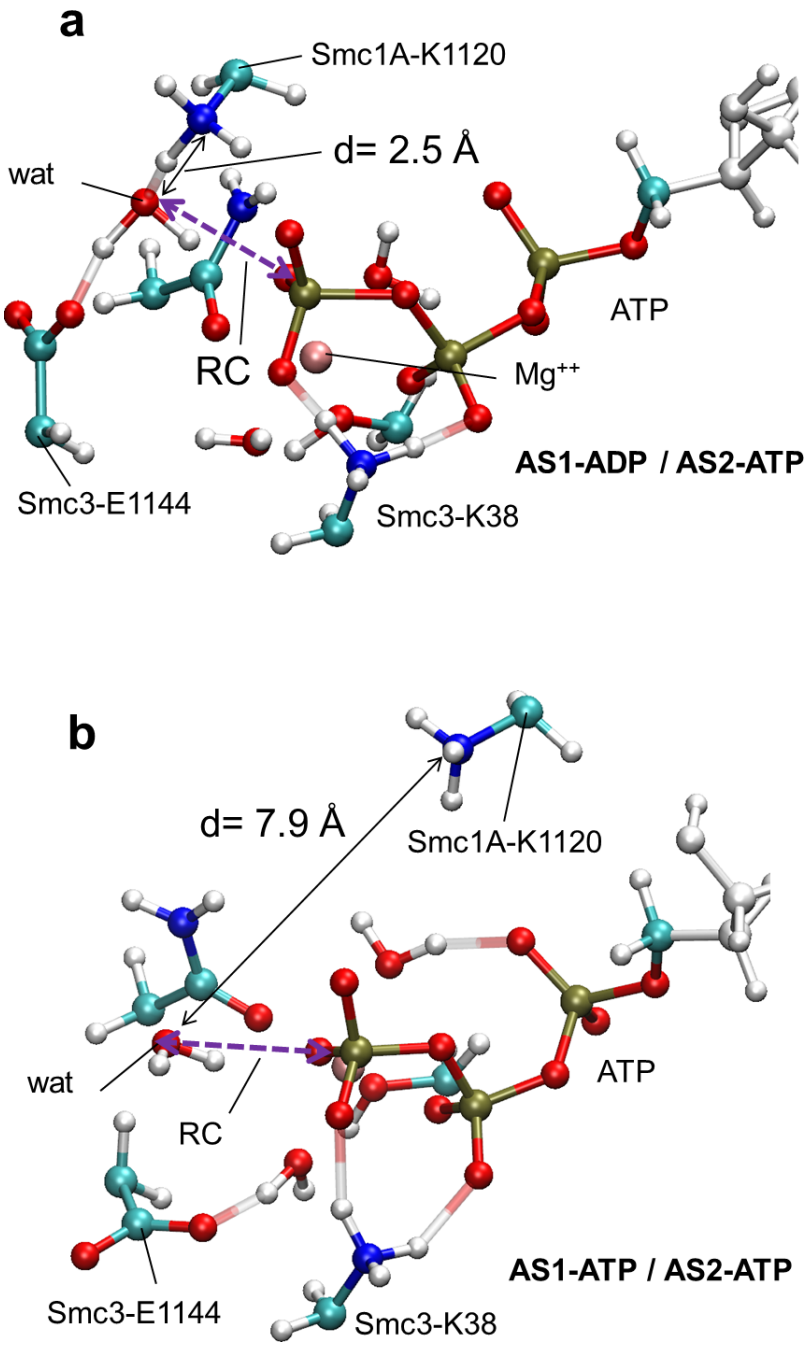


Figure 55: *QM region for AS2. Atoms belonging the QM region are depicted with colored ball and sticks. Light-gray ball and sticks conform the rest of the ATP molecule that is not contained in the QM region. The positions of the catalytic water (wat), Smc3-K38, Smc3-E1144 and Smc1A-K1120 residues, and the ATP molecule are indicated for both active (**a**) and inactive (**b**) AS2 configurations. The distance between the catalytic water and ϵ-amino group of the Smc1A-K1120 residue is indicated by a black arrow. The reaction coordinate (RC) is indicated by a purple arrow. Source: Marcos-Alcalde et al. (2017)[24]. Caption was adapted from the same source.*

catalytic water molecule and the γ-phosphate group of the ATP molecule (Fig. 55 a and appendix D). The strong electrostatic effects derived from such interactions were expected to produce two effects: water entrance and transition state stabilization. To evaluate this, ATP hydrolysis in AS2 was analyzed in both conditions (AS1-ADP/AS2-ATP and the control AS1-ATP/AS2-ATP) by QM/MM SMD with Fireball/AMBER. The final

structures of the 150 ns free MD stabilizations (Fig. 55) were used as initial structures for the QM/MM SMD simulations. The QM regions (Fig. 55) for both conditions were formed of the tri-phosphate moiety of the ATP in AS2, the magnesium ion, water molecules and side chains present in the coordination sphere of magnesium, the catalytic water molecule and the side chains of Smc3-K38, Smc3-E1144 and Smc1A-K1120. The MM region included the remaining atoms (i.e. the rest of the protein complex, counterions, solvent as well as AS1 atoms).

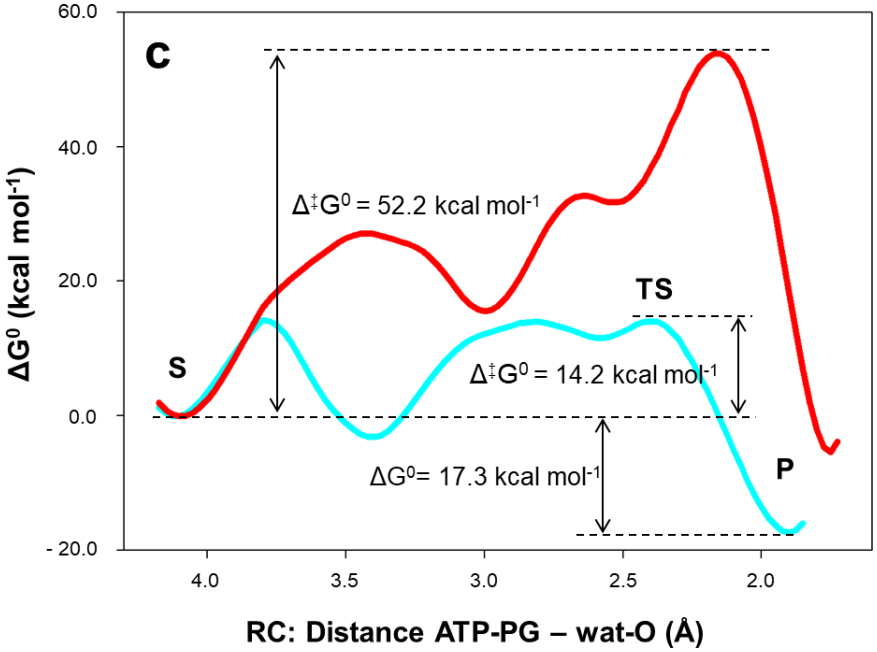


Figure 56: *ATP hydrolysis in AS1 facilitates ATP hydrolysis in AS2. A comparison of the free energy profiles of ATP hydrolysis in AS2 in while AS1 binds ADP (AS1-ADP/AS2-ATP; cyan) or ATP (AS1-ATP/AS2-ATP; red) is shown, indicating the location of the substrate (S), transition state (TS) and product (P). Activation free energy for both conditions ($\Delta^{\ddagger}G^{\circ}$) and free energy difference between substrate and product (ΔG°) in AS1-ADP/AS2-ATP condition are shown. Source: Marcos-Alcalde et al. (2017)*[24].

As a very noticeable difference between AS1-ADP/AS2-ATP and AS1-ATP/AS2-ATP was expected, instead of 3D free energy surface, a much less computationally expensive 2D free energy profile was calculated. The reaction coordinate (RC; purple arrow in Fig. 55) was defined to describe the bond to be formed, i.e. the distance between the oxygen atom from the catalytic water molecule and the phosphorus atom of the ATP γ-phosphate group. As expected, this approach was sensitive enough to detect a remarkable difference between the two conditions. In particular, dramatic reduction of the barriers associated with water entrance and transition state could be observed. The comparison of energy profiles shows that the activation of AS2 produced a total barrier reduction of 38 kcal mol^{-1}, leading to a final barrier value of 14.2 kcal mol^{-1} in AS1-ADP/AS2-ATP condition (Fig. 56). A detailed description of the observed key steps is available in (Fig. 58) and a video sequence of the reaction in AS1-ADP/AS2-ATP is shown in appendix E. This results support that ATP hydrolysis in AS1 leads to the entrance of Smc1A-K1120 in AS2, which strongly induces ATP hydrolysis. To evaluate the conservation of Smc1A-K1120 a MSA of the UniProtKB canonical sequences for proteins coded by orthologous genes to human SMC1A, SMC3, SMC2 and SMC4 was obtained (Fig. 57). This MSA confirmed that K1120 is conserved in SMC1A orthologs from human to yeast but, quite

interestingly, that it is present in SMC4 genes too, which are the molecular equivalent to SMC1A in condensin. This result not only supports Smc1A-K1120 to be playing a strong functional role in SMC1A but also leaves the open question of whether this molecular mechanism could be generalized for both cohesin and condensin.

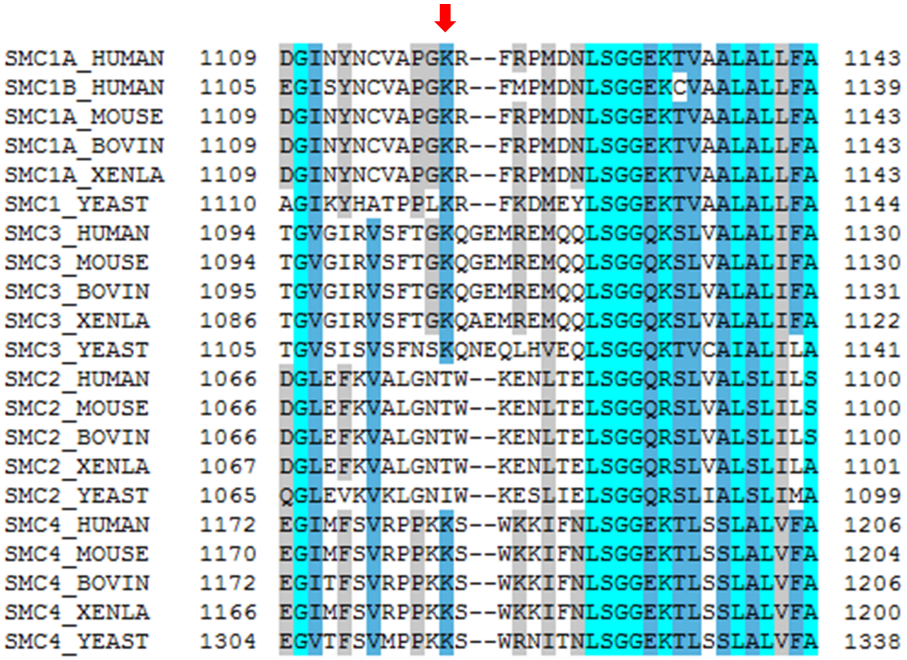


Figure 57: *Multiple sequence alignment of several proteins homologous to human Smc1A in the area surrounding residue K1120. The sequences represented are: human Smc1A (SMC1A_HUMAN), Smc1B (SMC1B_HUMAN), Smc3 (SMC3_HUMAN), Smc2 (SMC2_HUMAN) and Smc4 (SMC4_HUMAN); Mus musculus Smc1A (SMC1A_MOUSE), Smc3 (SMC3_MOUSE), Smc2 (SMC2_MOUSE) and Smc4 (SMC4_MOUSE); Bos taurus Smc1A (SMC1A_BOVIN), Smc3 (SMC3_BOVIN), Smc2 (SMC2_BOVIN) and Smc4 (SMC4_BOVIN); Xenopus laevis Smc1A (SMC1A_XENLA), Smc3 (SMC3_XENLA), Smc2 (SMC2_XENLA) and Smc4 (SMC4_XENLA); and Saccharomyces cerevisiae Smc1 (SMC1_YEAST), Smc3 (SMC3_YEAST), Smc2 (SMC2_YEAST) and Smc4 (SMC4_YEAST). The residues are colored according to conservation (BLOSUM62 score). The position of human Smc1A-K1120 is indicated by an arrow. Source: Marcos-Alcalde et al. (2017)[24]. Caption was directly transcribed from the same source.*

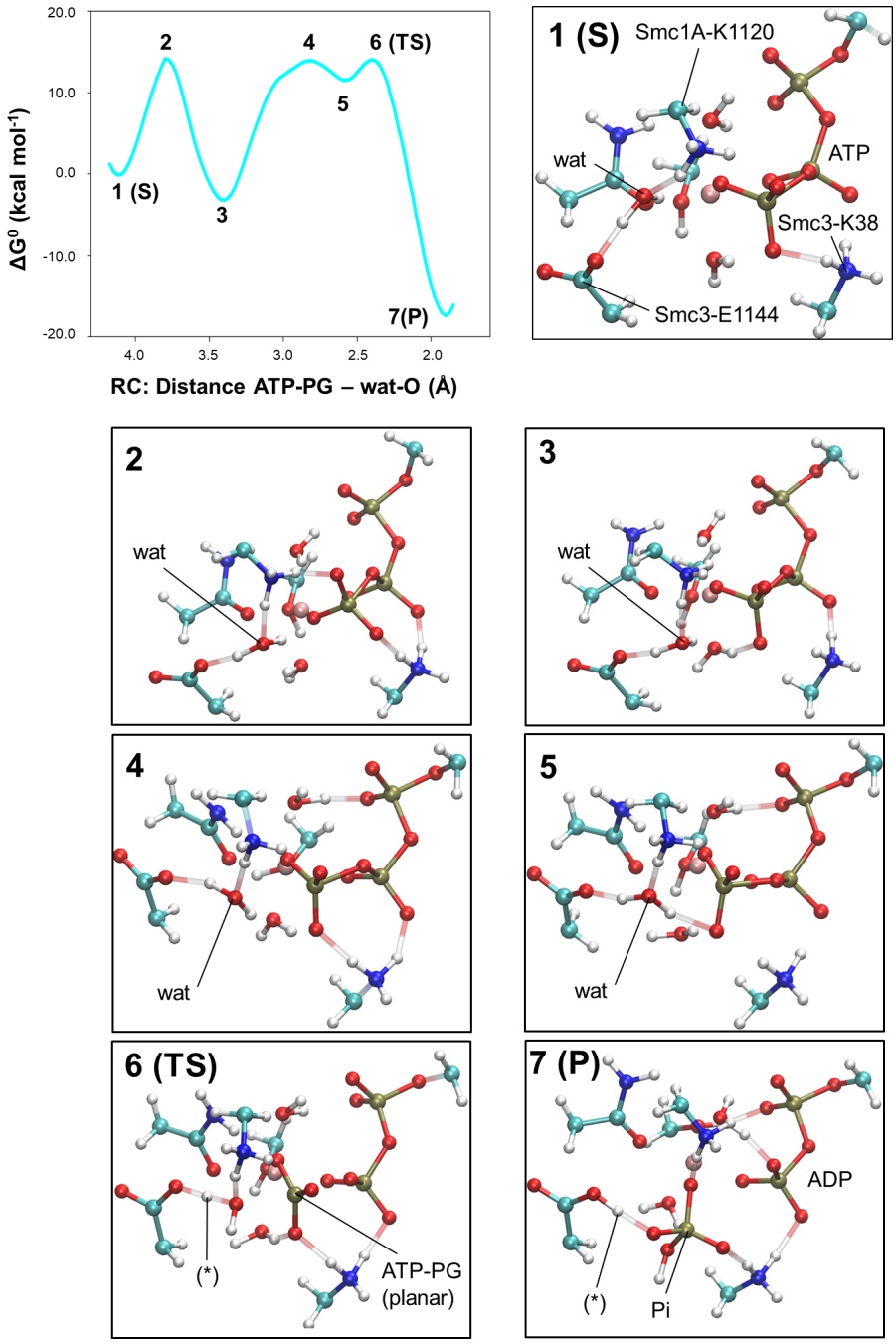


Figure 58: *Points of interest in the energy profile of the ATP hydrolysis at AS2 in its active form (AS1-ADP/AS2-ATP). Points of interest (1-7) of the free energy profile are indicated and a reference structure is shown for each one. The energy profile can be divided into 3 steps: the first step (panels 1 to 3) is characterized by the entrance of the catalytic water molecule in the active site; the second (panels 3 to 5) corresponds to the stabilization of the catalytic water molecule prior to the transition state by triple hydrogen bonding with Smc3-E1144, Smc1A-K1120 and γ-phosphate; the third (panels 6 and 7) correlates to the transition state, thus leading to the product structure. The leaving inorganic phosphate group is formed and Smc3-E1144 transfers the catalytic water proton (*) to the leaving group. Source: Marcos-Alcalde et al. (2017)[24]. Caption was directly transcribed from the same source.*

3.2.2.3 ATP hydrolysis facilitates separation of the ATPase heads

Our results support the hypothesis that Rad21-Cter binding facilitates ATP hydrolysis in AS1, which induces hydrolysis in AS2. Furthermore, biochemical literature describes that ATP binding and hydrolysis are required for both DNA loading[135–138] and unloading[132,139,140] and mechanistic models of these two processes assume that the ATPase head domain has to separate to let DNA pass through[128,132,137–140,165,166]. Consequently, we evaluated the effect the exchange of ATP with ADP (as a result of the aforementioned ATPase activity) has over head heterodimer stability.

In order to describe this configuration, two equivalent models of the Smc1A-head Smc3-head Rad21-Cter complex were generated binding two molecules of either ATP or ADP. The ATP binding (AS1-ATP/AS2-ATP) model represents the complex structure prior to any ATP hydrolysis event, while the ADP binding (AS1-ADP/AS2-ADP) model represents the complex structure after hydrolysis in both active sites. Both models were stabilized over 150 ns long free MD simulations from which individual structures were extracted every 4 ns, from 104 ns to 120 ns, yielding 5 extracted structures per condition. As these were to be used as initial structures in the succeeding SMD simulations, they were extracted from a stable region ensuring that a period of stability longer than the SMD simulation times (13 ns) came after the extraction times (30 ns in the shortest case, i.e. 120 ns extraction). Ten SMD simulations, five per condition, were generated starting from the extracted structures, forcing the centers of mass of Smc1-head and Smc3-head to separate from each other 32.5 Å along 13.0 ns (Fig. 59), measuring the accumulated work along each trajectory. Using Jarzynski's equality (Eq. 1.13) we estimated the ΔG° difference between the closed and opened conformations of the complex models (quasi-equilibrium states) in both conditions (AS1-ATP/AS2-ATP and AS1-ADP/AS2-ADP) from the accumulated work data obtained along an ensemble of SMD trajectories simulating the heterodimer separation (non-equilibrium transitions between quasi-equilibrium states). To better reach the required quasi-equilibrium state in both closed and opened conformations, the center of mass distance separation was kept constant for 0.1 ns during the start and end points of each SMD trajectory.

The steepest slope of the free energy profile (Fig. 59 b) and, consistently, the largest force peak along each trajectory (Fig. 59 a), both take place during the first ∼7.5 Å of separation and are the most distinctive feature between the trajectory ensembles representing each condition. Interestingly, during this separation period, all the interactions between residues in both sides of each active site were broken. Unsurprisingly, the most noticeable difference to be observed over the free energy eestimations arises at that region, being the force peak associated with active site disruption the largest contribution to the 24.8 kcal mol^{-1} ΔG° difference between AS1-ATP/AS2-ATP and AS1-ADP/AS2-ADP conditions (Fig. 59 c). It is noteworthy that 24.8 kcal mol^{-1} is approximately 81% of the average ΔG° associated with the hydrolysis of two molecules of ATP in human muscle in resting conditions: ΔG° 30.6 kcal mol^{-1} [167]. This could be hinting that cohesin ATPase head would be highly efficient from an energetic point. Together these results support the hypothesis that the ATP hydrolysis at both AS1 and AS2 significantly facilitates ATPase head complex opening.

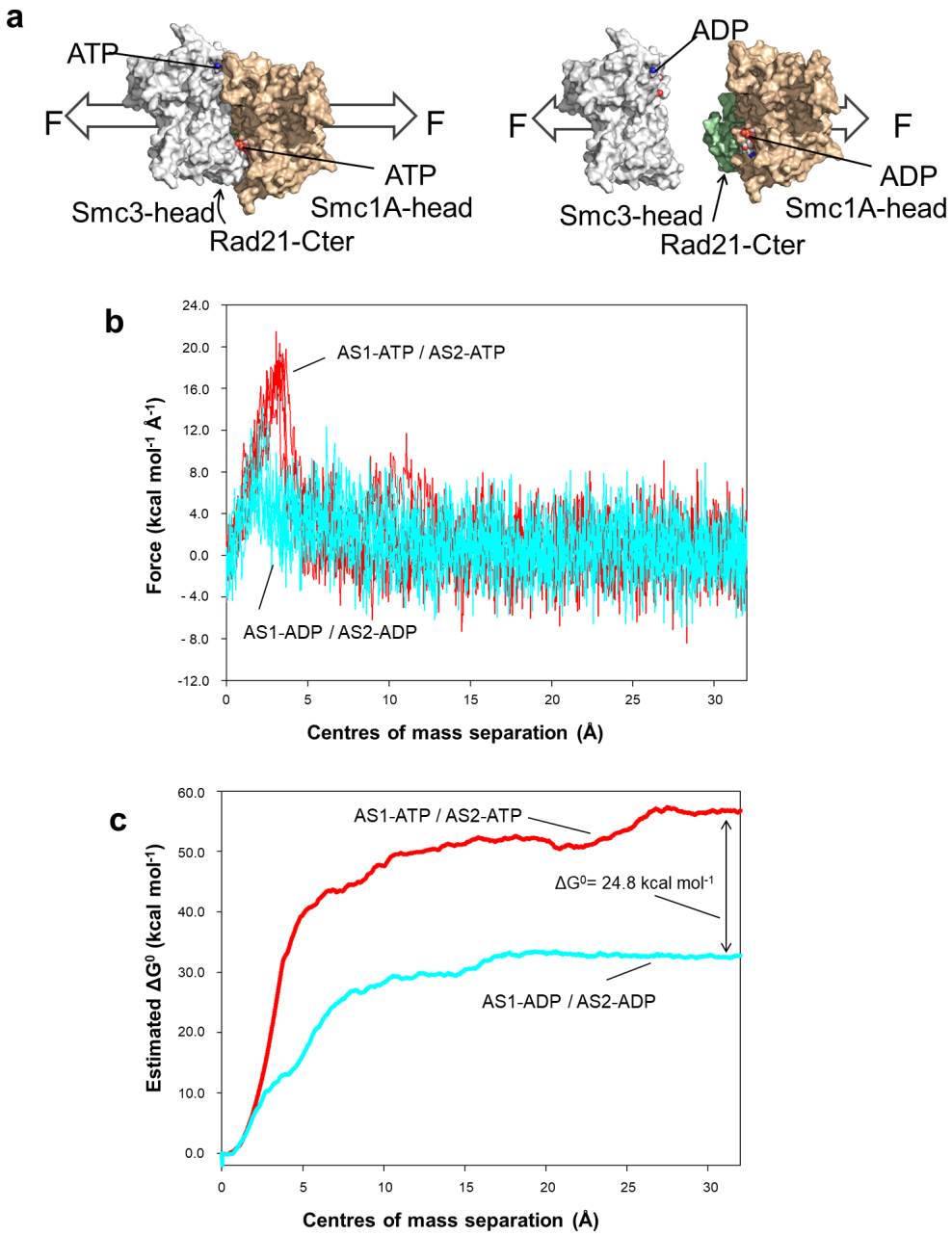


Figure 59: *ATP hydrolysis at AS1 and AS2 facilitates head separation. (**a**) Schematic overview of the head separation induced by SMD simulations. The Smc1A-head (brown), Smc3-head (gray) and Rad21-Cter (green) domains are shown and the nucleotide (ATP or ADP) locations in both active sites are indicated. Force (F) direction is marked by white arrows. (**b**) Forces exerted in SMD trajectories over the separation between the centers of mass of Smc1A-head and Smc3-head domains. Points from all trajectories for the AS1-ATP/AS2-ATP condition (red) and for AS1-ADP/AS2-ADP condition (cyan) are shown. (**c**) Estimated free energy difference (kcal mol^{-1}) over the separation between the centers of mass of the Smc1A-head and Smc3-head domains computed using Jarzynski's equality over 5 SMD trajectories for each condition. Source: Marcos-Alcalde et al. (2017)[24]. Caption was directly transcribed from the same source.*

3.2.2.4 Pathogenic variants and mutants with an associated phenotypic effect

The cohesin ATPase head complex dynamic models we generated in order to appraise the ATP hydrolysis dynamics and its possible influence over the complex stability also provided an atomistic framework quite suited for rationalizing mutations linked with pathologies and phenotypic variations. Thirteen human pathogenic variants (Table 3 and Fig. 60 b,d,e) and three yeast non-neutral (i.e. phenotype-changing) mutations (Fig. 60 f) were analyzed by means of this framework. These variants, all of which are predicted to affect ATP hydrolysis in either AS1 or AS2, were arbitrarily grouped into four clusters for clarity purposes. The first cluster (depicted in green in figure 60) was comprised of the residues Smc1A-N34, Smc1A-R57, Smc3-G1118 and Smc3-Q1119; and, according to our models, mutations altering those residues are expected to affect ATP positioning and ATP hydrolysis progress. Smc1A-N34 stabilizes the entrance of the catalytic water molecule (Fig. 51 c "2") and the planar transition state (Fig. 51 c "4(TS)"). Smc1A-R57 interacts with α-phosphate group of the ATP molecule, stabilizing its position in the active site. Smc3-G1118 and Smc3-Q1119, located in the immediate proximity of Smc1A-N34 are expected to keep Smc1A-N34 in the right place and orientation. Mutations affecting the residues grouped in cluster one have been related to endometroid carcinoma (Smc1A-N34 and Smc1A-R57)[168] and acute myeloid leukaemia (Smc3-G1118 and Smc3-Q1119)[18,168].

The second cluster (depicted in yellow in figure 60) was comprised of the residues Smc1A-N1166, Smc3-D1143, Smc3-Q1147 and Smc3-A1148. Smc1A-N1166T and Smc3-Q1147E variants have been found in CdLS patients[9,142,171] whereas mutations affecting residues Smc3-D1143 and Smc3-A1148 have been related to acute myeloid leukaemia[18,168] and colorectal cancer[168,172] respectively. In a previous work[9], in which our group collaborated, CdLS related Smc3-Q1147E mutation was analyzed via homology modeling, proposing that, given Smc3-Q1147 location in that model, the mutation could potentially be affecting interactions with Smc1A that could affect dimer stability and/or ATPase activity in AS2. Interestingly, in our dynamic framework, when AS2 activation by Smc1A-K1120 entrance was described, Smc3-Q1147 was found to be in close proximity to the incoming α-amino group of Smc1A-K1120. Mutation of Smc3-Q1147 to a negatively charged glutamic acid

Protein	Mutation	Disease	Location	References
Smc1A	N34S	Endometroid carcinoma	Active site 1	168
Smc1A	R57W	Endometroid carcinoma	Active site 1	168
Smc1A	V58_R62del	Cornelia de Lange Syndrome	Putative binding to DNA	122, 142, 143, 169
Smc1A	R1090C	Melanoma	Active site 2(activation)	17, 168
Smc1A	F1122L	Cornelia de Lange Syndrome	Active site 2 (activation)	122, 143, 170
Smc1A	R1123W	Cornelia de Lange Syndrome	Active site 2 (activation)	143, 144
Smc1A	N1166T	Cornelia de Lange Syndrome	Active site 2	142
Smc3	H55Y	Colorectal cancer	Putative binding to DNA	19, 168
Smc3	G1118V	Acute myeloid leukaemia	Active site 1	18, 168
Smc3	Q1119K	Acute myeloid leukaemia	Active site 1	18, 168
Smc3	D1143H	Acute myeloid leukaemia	Active site 2	18, 168
Smc3	Q1147E	Cornelia de Lange Syndrome	Active site 2(activation)	9, 171
Smc3	A1148T	Colorectal cancer	Active site 2	168, 172

Table 3: *Human pathogenic variants.*

could be expected to induce strong interaction with the positively charge α-amino group, probably altering the interactions in AS2 that lead to ATPase activity enhancement. To evaluate this possibility, analogously to the simulations in which AS1-ATP/AS2-ATP and AS1-ADP/AS2-ATP conditions were compared, a new model of the Smc1A-head Smc3-head Rad21-Cter complex, in presence of ADP in AS1 and ATP in AS2 and having Smc3-Q1147 replaced by glutamic acid (Smc3-Q1147E/AS1-ADP/AS2-ATP condition), was subjected to 150 ns of free MD. In the new trajectory, the distance between Smc1A-K1120 and the catalytic water molecule of AS2 (Fi.g 60 c) showed an intermediate behavior to those shown by the trajectories performed in AS1-ATP/AS2-ATP and AS1-ADP/AS2-ATP conditions. Compared to AS1-ADP/AS2-ATP condition, in which AS2 showed an active arrangement during 13.45% of the total 150 ns, in Smc3-Q1147E/AS1-ADP/AS2-ATP AS2 only showed an active arrangement during 0.03% of the 150 ns. This suggests a greatly reduced, yet not completely abolished, ATPase activity at the AS2 of the cohesin molecules of the CdLS patient with such mutation. To us this finding was very exciting as, to the best of our knowledge, this was the first time for a CdLS related variant to be assigned a specific functional role in the cohesin complex dynamics.

The third cluster (depicted in magenta in Fig. 60) was comprised of the residues Smc1A-R1090, Smc1A-F1122 and Smc1A-R1123. Smc1A-F1122L and Smc1A-R1123W variants have been found in CdLS patients[123,143,144,170] and the Smc1A-R1090C variant has been related to melanoma[17,168]. The most noticeable feature of these three residues is that their positions are closely related to the movement of Smc1A-K1120 and all three mutations can potentially disrupt the correct orientation of Smc1A-K1120 towards AS2.

The fourth cluster (depicted in pink in Fig. 60) was comprised of Smc1A-L1128, Smc1A-G1131 and Smc1A-D1163. These residues are human Smc1A orthologous positions to yeast Smc1 residues (Smc1-L1129, Smc1-G1132 and Smc1-D1164) that, when mutated, can bypass the need for Eco1[139] (Fig. 61). Smc1A-G1131 and Smc1A-D1163 are close to the γ-phosphate group of the ATP molecule in AS2 and, thus, mutations can be expected to alter ATP hydrolysis dynamics in this active site. However, offering a possible mechanistic explanation for yeast Smc1-L1129V (human Smc1A-L1128) mutation affecting the rate of DNA release of cohesin poses a much more challenging task as the orthologous mutation in yeast Smc3 (Smc3-L1126V) does not. Interestingly, in our simulation, Smc1A-L1128 hydrophobic side chain stabilized the hydrocarbon chain of Smc1A-K1120, therefore forcing its correct orientation and subsequent entrance towards AS2 ATP γ-phosphate group. Leucine to valine substitution, otherwise a non-drastic mutation, reduces the length of the non-polar side chain of leucine, thus shrinking the surface along which the hydrophobic interaction stabilizes Smc1A-K1120 side chain orientation. If true, this molecular mechanism could explain why yeast Smc3 orthologous mutation fails to reproduce Smc1-L1129V phenotype as Smc3-L1115 (human ortholog to Smc3-L1126V), similarly to Smc1A-L1128, is close to the active site (AS1 in this case) but, due to its location, would not play a direct role in ATP hydrolysis. On contrast, while Smc1A-L1128 would neither play a direct role in ATP hydrolysis, it would be required for the activation of AS2 by Smc1A-K1120.

Additionally to those residues that can be linked with pathologies and phenotypic variations, we paid special attention to residues related to acetylation-regulated DNA binding, for they are crucial for the ultimate understanding of the cohesin function. These can be located both in the coiled-coils[173], almost completely removed in our model, and in the inner side of the ATPase head domains[128,132,140,174]. Probably due to the absence of

forces exerted by the unmodeled region of the cohesin ring (coiled-coils dimerized by the hinge region), during the Rad21-Cter activation simulations, after the first 40 ns of free MD simulations performed with our models, the relative angle between Smc1A-head and Smc3-head monomers grew wider (Fig. 62 a) resembling structures of Rad50 (a SMC family member) binding a DNA double strand[175–177]. The root mean square deviation (RMSD) of Smc1A-head and Smc3-head were measured for the whole 120 ns long free MD simulation to which the Smc1A-head Smc3-head Rad21-Cter complex was subjected (Fig. 62 b), evaluating that the internal structure of each monomer remained stable. The RMSD values were constantly bellow 3.0 Å, despite change in the internal angle. During this free MD simulation a number of positively charged residues relocated towards the inner surface of the ATPase head dimer (Fig. 62): Smc1A-K59, Smc1A-R62, Smc1A-K149, Smc3-H55, Smc3-R61, Smc3-K105, Smc3-K106 and Smc3-K157. Interestingly, Smc3-K105 and Smc3-K106 are described to be involved in acetylation-regulated DNA binding[128,132,140,174], the deletion of Smc1A-K59 and Smc1A-R62 has been related to CdLS[122,142,143,169] and Smc3-H55Y mutation has been associated with colorectal cancer[19,168]. Remarkably, after relocation, the positions of all these positive residues becomes fully compatible with the putative position of a double stranded DNA molecule over the inner surface of the ATPase head complex (Fig. 62), similarly to the Rad50 structures binding double stranded DNA. This observation may suggest that our model might be mimicking, not only the binding of Rad21-Cter, but also a geometry somehow similar to that of the DNA bound structure, resembling to an extent the starting event of the ATPase-dependent opening of the cohesin head.

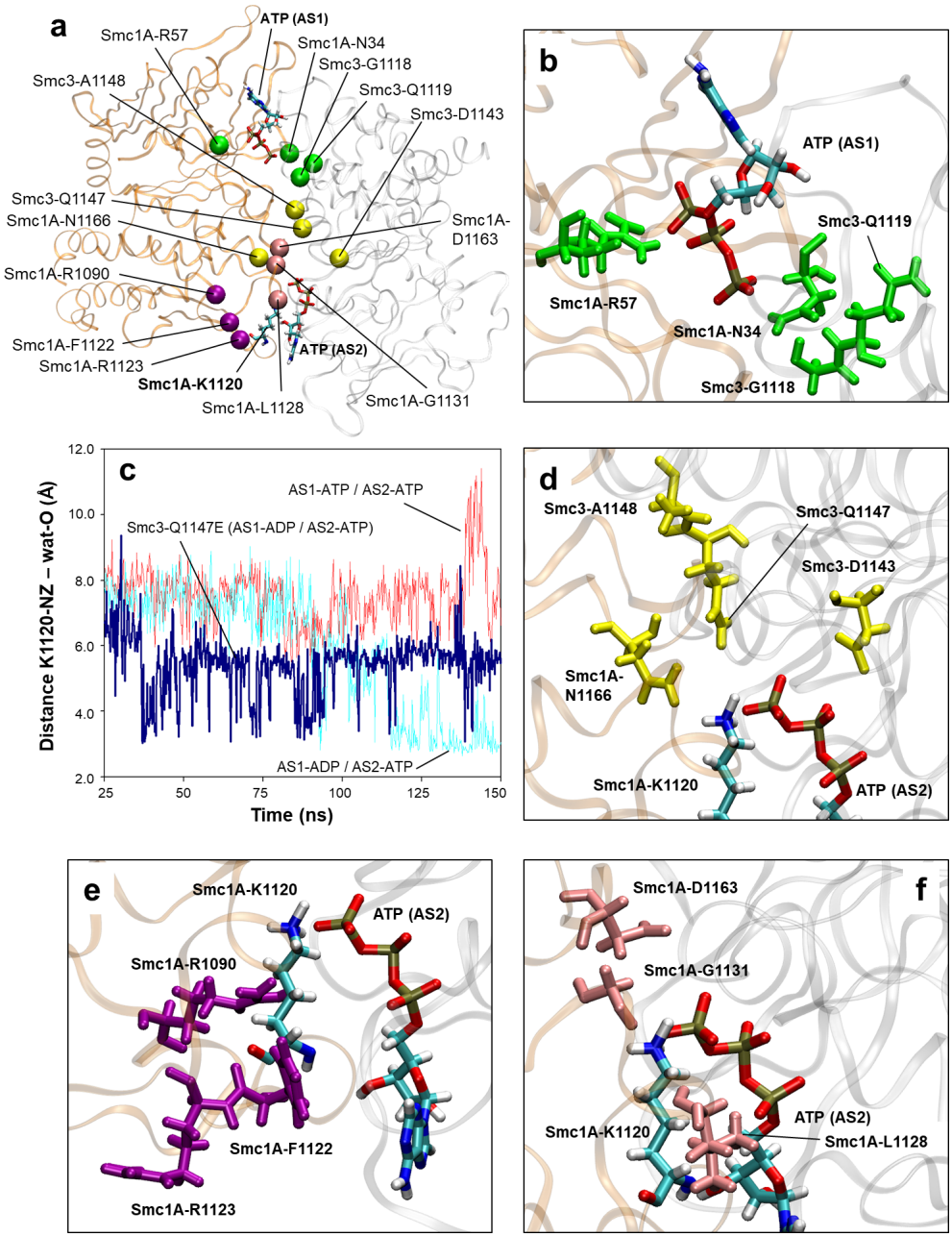


Figure 60: *Pathogenic variants and non-neutral mutations. (**a**) Location of the Ca atoms of residues of interest in the neighborhood of AS1 and AS2. Disease-related variants are shown in green (those affecting AS1), purple (those affecting AS2 activation via Smc1A-K1120 rearrangement) and yellow (those affecting AS2 directly). Residues equivalent to those affected by mutations that bypass the need for Eco1 in yeast are shown in pink. The Smc1A-K1120 residue and both ATP molecules are shown. (**b**) Location of the variants affecting AS1. Residues are depicted in green. (**c**) Evolution of the distance between the oxygen atom of the catalytic water in AS2 and the ϵ-amino group of the Smc1A-K1120 residue (distance K1120-NZ - wat-O). Distances obtained with wild-type Smc3 prior (red) and after (cyan) ATP hydrolysis at AS1, and distances obtained with the Smc3-N1147E mutant after ATP hydrolysis at AS1 (blue) are shown. (**d**) Location of the variants directly affecting AS2. Residues are depicted in yellow. (**e**) Location of the variants affecting AS2 activation via K1120 rearrangement. Residues are depicted in magenta. The location of K1120 is indicated. (**f**) Location of the residues equivalent to those affected by mutations that bypass the need for Eco1 in yeast. Residues are depicted in pink. The location of K1120 is indicated. Source: Marcos-Alcalde et al. (2017)[24]. Caption was directly transcribed from the same source.*

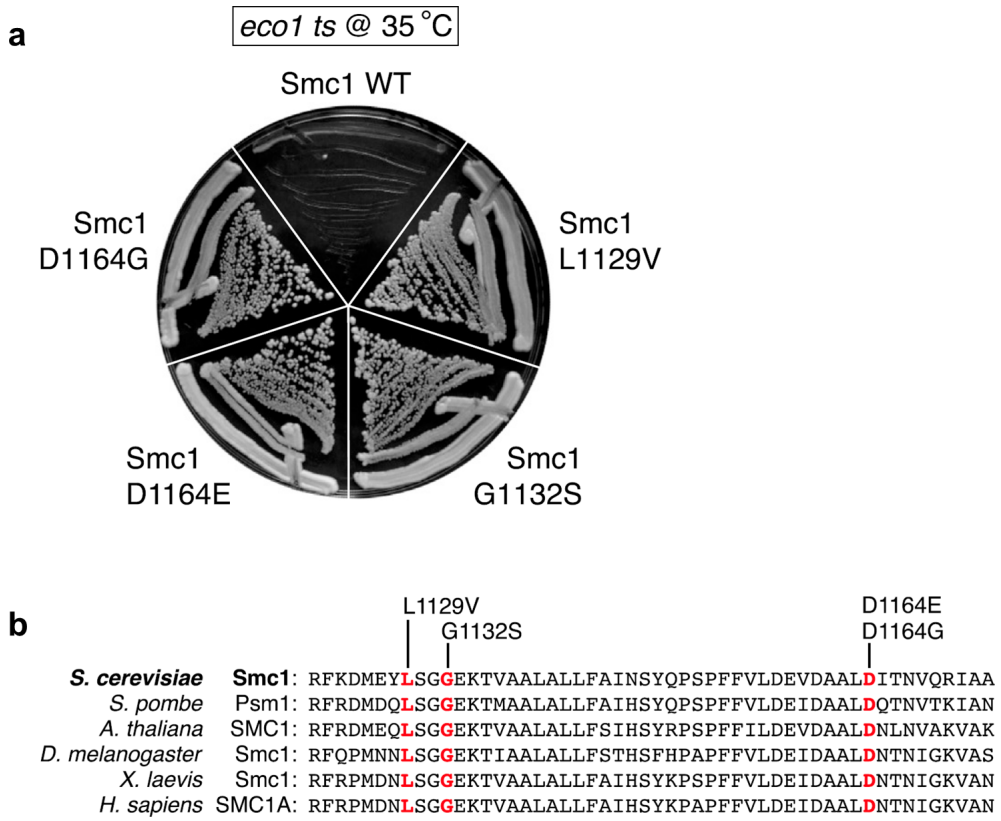


Figure 61: *Mutations that bypass the need for Eco1 in yeast. The lethal phenotype observed when a yeast strain carrying a temperature-sensitive Eco1 allele is grown at 35° C is rescued by Smc1-L1129V, Smc1-G1132S, Smc1-D1164E and Smc1-D1164G mutations (**a**). Smc1-L1129, Smc1-G1132, Smc1-D1164 are conserved residues in Smc1 orthologs from yeast to human (**b**). Source: Elbatsh et al. (2016)*[139].

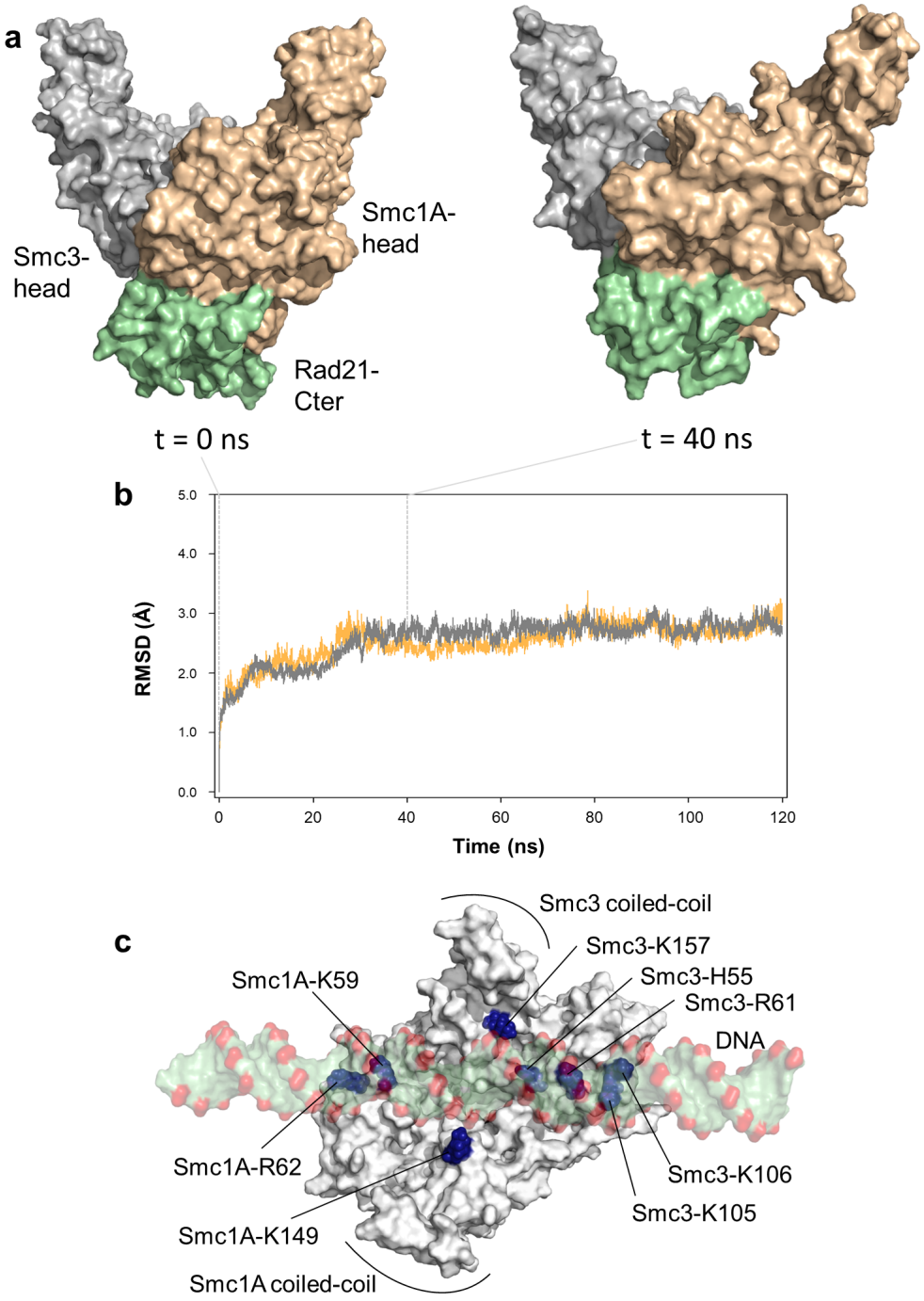


Figure 62: *AS1-ATP/AS2-ATP trajectory spontaneously exhibits a conformation compatible with DNA binding. (**a**) Structural evolution of the Smc1A-head, Smc3-head and Rad21-Cter complex in AS1-ATP/AS2-ATP condition. The equilibrated model structure that was used as the initial structure in MD simulations (t = 0 ns) and the conformation it adopted after 40 ns of free MD (t = 40 ns) are shown. (**b**) RMSD values measured over the unrestricted 120 ns MD trajectory of the complex illustrated in a is shown, indicating the simulation times that correspond to the structures depicted in (**a**). Note that RMSD remains stable after ∼ 30 ns. (**c**) Position of positively charged residues in the upper surface of the head complex after 40 ns of MD. The putative position of a DNA molecule, in the equivalent position as the one co-crystallized with Rad50 head domain (PDB code: 5DNY), is indicated. Modified from Marcos-Alcalde et al. (2017)[24]. Caption was adapted from the same source.*

3.2.3 Discussion

Cohesin complex plays a fundamental role in faithful genome segregation, a universal requirement among all living beings. This complex also participates in many other crucial processes such as DNA repair, chromatin organization and transcription regulation[120–123]. Subtle alterations in this complex can lead to developmental syndromes, such as Roberts Syndrome, Warsaw Breakage Syndrome, CAID syndrome or CHOPS syndrome and, the most prevalent of them all, Cornelia de Lange Syndrome[8–13,142–144] as well as several types of cancer[14–22]. Given the fundamental character of cohesin any major alterations on this complex probably will make development inviable, leading to lethal phenotypes that cannot be observed. This fact, combined with the extraordinary size of SMC proteins coiled-coil regions, has made mechanistic explorations of these protein complexes highly challenging for biochemists as well as structural and computational biologists. The resolution of structures of SMC proteins, even if truncated, has proven to be a powerful resource for the proposal of new mechanistic hypotheses and biochemical experiments[7]. Making use of the structural information available we built various homology models of the human Smc1A-head Smc3-head Rad21-Cter complex reproducing different events associated with the ATPase activity of the cohesin ATPase head. These models were subjected to a series of simulation procedures letting these structures relax in a thermalized environment (free MD), comparatively simulating ATP hydrolysis in different conditions (QM/MM SMD) and simulating the separation of SMC proteins in the complex before and after ATP hydrolysis (SMD) with the aim of reproducing the series of events that lead to the DNA passing through the complex (Fig. 63).

The first event to be studied was the binding of Rad21-Cter the Smc1A-head Smc3-head complex, investigating the molecular mechanism leading to the experimentally observed ATP hydrolysis induction of the Smc1A-head Smc3-head dimer by Rad21-Cter binding to Smc1A-head[141]. To such end, a homology model of the human ATPase head complex was obtained. The X-ray structure of a forced yeast Smc1 (human Smc1A ortholog) homodimer bound to the C-terminal domain of Scc1 (human Rad21 ortholog)[133] was used as the main scaffold, as it contained detailed information about the interface between SMC subunits. A yeast Smc3 (human Smc3 ortholog) monomeric structure[131] was aligned with one of the Smc1 monomers generating an Smc1 Smc3 complex with Scc1 bound to Smc1. Lastly, to refine the structure, a high resolution *Pyrococcus furiosus* Smc homodimer structure[178] was used to determine the position of not clearly located residues surrounding the active sites as well as to incorporate the crystallographic water molecules present in this structure into the final model. Using this complex scaffold a homology model of the Smc1A-head Smc3-head Rad21-Cter complex, with crystallographic water molecules and binding ATP in both active sites was obtained. In order to reproduce different conditions this first model was modified several times, either removing Rad21-Cter, changing the nucleotides bound to the active sites in various configurations or modeling a CdLS related variant (Smc3-Q1147E). The first model (in presence of Rad21-Cter and binding ATP in AS1 and AS2) underwent a spontaneous rearrangement, probably due to the absence forces derived from the unmodeled coiled coil and hinge regions, when relaxed through free MD simulations, exposing a group of eight positively charged residues towards the inner surface of the complex (Fig. 62 c). The resulting conformation was compatible with a DNA-bound conformation of the Smc1A-head Smc3-head Rad21-Cter complex (Fig. 62 c). It is noteworthy that this group of spontaneously exposed amino acids contains two lysine residues that had already been proposed to be involved in acetylation-regulated

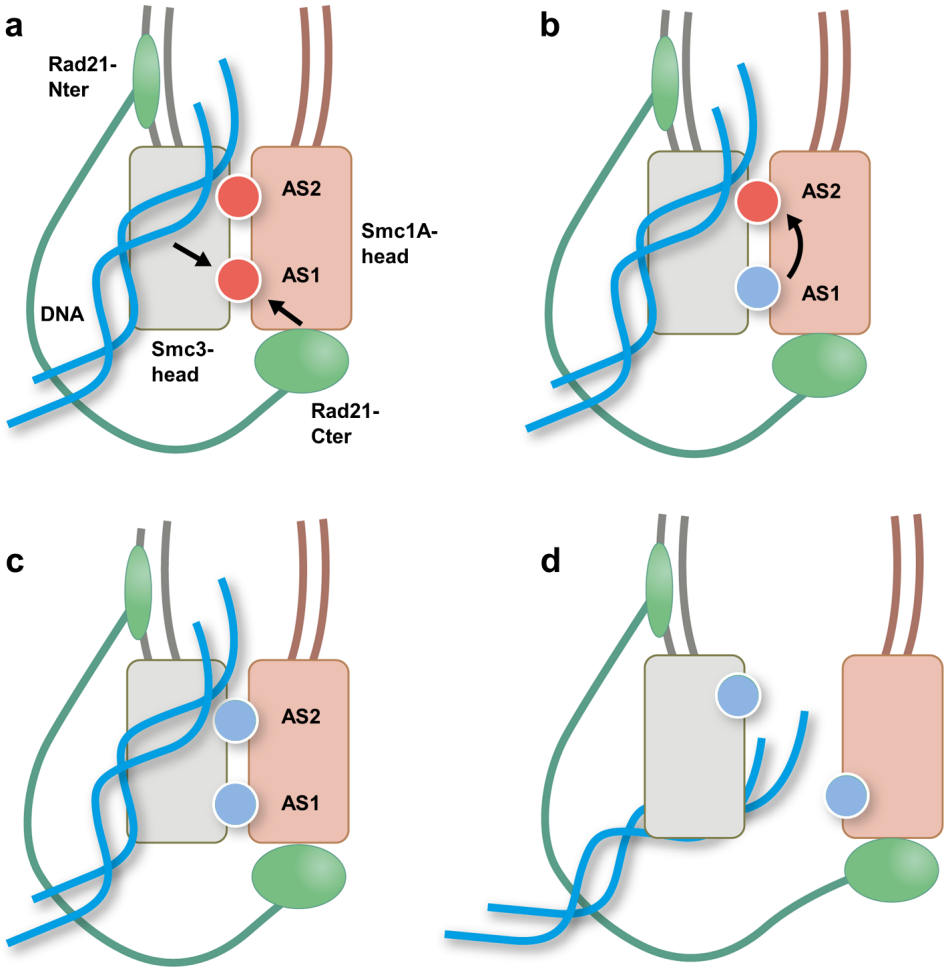


Figure 63: *Schematic model for ATP hydrolysis-driven head opening. (**a**) The Rad21-Cter domain binding to the Smc1A-head domain allows hydrolysis at AS1. (**b**) ATP hydrolysis at AS1 induces AS2 activation via Smc1A-K1120 rearrangement. (**c**) ATP hydrolysis takes place at AS2. (**d**) ATP hydrolysis at both active sites facilitates the separation of the ATPase head domains. Source: Marcos-Alcalde et al. (2017)[24]. Caption was directly transcribed from the same source.*

DNA binding[128,132,140,174], as well as other three residues the mutation of which has been related to CdLS[142–144] and cancer[19,168]. The fact five out of eight residues exposed are either already described as involved with DNA binding or related with human diseases reinforces the non-spurious character of this rearrangement.

After stabilizing with free MD the model in presence and absence of Rad21-Cter, ATP hydrolysis in both conditions was simulated obtaining a much favorable reaction path in presence Rad21-Cter (reducing the free energy barrier by 14.0 kcal mol^{-1}) and obtaining atomistic description of the key steps in the reaction, such as the catalytic water entrance and the planar transition state stabilization (Fig. 51 c "4(TS)" and appendix C).

In many ATP-binding-cassette (ABC) ATPases the hydrolysis of ATP at one active site can stimulate ATP hydrolysis at the other[179]. As Smc1A-head and Smc3-head constitute a heterodimeric ABC ATPase domain we wondered whether this allosteric coupling between active sites could be conserved in cohesin. Free MD stabilization of the Smc1A-head Smc3-head Rad21-Cter model binding ADP to AS1 and ATP to AS2 revealed that Smc1A-K1120, a previously unrelated residue, spontaneously relocated towards AS2, in such way

that the α-amino group eventually formed stable hydrogen bonds with both the oxygen atom of the catalytic water molecule and the γ-phosphate group of the ATP molecule (figures 54 and 55 and appendix D). As all of the critical events in the reaction imply partial and net negative charges being close to each other, the entrance of a positive net charge in that region of the active site was expected to induce a strong electrostatic stabilization and thus a dramatic reduction of the free energy barrier. This was tested with a simpler, yet sensitive enough, simulation of the reaction which confirmed an extensive barrier reduction after Smc1A-K1120 entrance in AS2 and offered atomistic descriptions of the key steps in the reaction (figure 58 and appendix E). Therefore, our model predicts that the cohesin head complex exhibits allosteric coupling between AS1 and AS2, similarly to other ABC ATPases[179]. Unfortunately we have not been able to characterize the "driving force" of Smc1A-K1120 movement yet. However, it is interesting to note that Smc1A-K1120 is conserved in all Smc1A and Smc1B sequences as well as in the majority of the SMC proteins (Fig. 57). A striking exception is Smc2, the protein that plays the equivalent role of Smc3 in condensin[121,127], in which lysine residue is replaced by threonine (Smc2-T1077 in human) or isoleucine in different organisms (Fig. 57), while in Smc4, the protein that plays the equivalent role of Smc1A in condensin, lysine residue (Smc4-K1183 in human) is conserved (Fig. 57). If a similar allosteric coupling between AS1 and AS2 is to be assumed in both cohesin and condensin heterodimers, Smc2-T1077, unlike Smc4-K1183, would not be necessary for the intramolecular activation to occur, which is consistent with the observed conservation profile of both residues. It might also be notorious that most of the variants related with CdLS and cancer roughly located in the putative pathway that would connect AS1 and AS2 (Fig. 60).

After having simulated ATP hydrolysis in AS1 and AS2, the subsequent head separation (Fig. 63 d) was evaluated, comparing pre-hydrolysis and post-hydrolysis conditions in order to evaluate the effect ATP hydrolysis could have over the stability of the cohesin ATPase head complex. After free MD stabilization of Smc1A-head Smc3-head Rad21-Cter model binding either ATP or ADP in both active sites, five separations were simulated for each condition via SMD starting from different points. Making use of Jarzynski's equality[84] an estimate of ΔG° was reconstructed from the SMD trajectories. The results (Fig. 59) suggest that ATP hydrolysis in both active sites is a major regulator of the complex stability and that this mechanism is highly efficient from an energetic point of view. Together, these observations support the hypothesis that ATP hydrolysis is followed by the opening of the ATPase head complex (Fig. 63 d). This is in agreement with recent x-ray diffraction structures of bacterial Smc homodimer[180] as well as data derived from a yeast Smc1-head Smc3-head heterodimer modeled by direct crystal structure alignment[165].

As previously commented, one of the most significant advantages of obtaining atomistic descriptions of biological processes is the ability to assign degrees of relevance to certain atoms or molecules, providing new experimental proposals as well as a certain degree of rationalization to previous experimental results. The atomistic framework developed in this thesis helped to rationalize several pathogenic variants related with cancer and CdLS (table 3 and figure 60). Most notably, the predicted allosteric coupling between AS1 and AS2 offered a possible mechanistic explanation for the Smc3-Q1147E variant, present in a CdLS patient[9,142,171]. As far as we know, this was the first time a CdLS variant had been mechanistically explained in detail. In our framework this variant was predicted to partially interfere in the activation of AS2 after ATP hydrolysis in AS1,

which can be expected to produce an impaired, yet not completely disrupted, behavior of the cohesin complex (Fig. 58). This effect is compatible with the pathogenic phenotype observed in the patient, who was phenotypically classified as "moderate" [169]. As previously commented, a more severe effect resulting in a complete loss of cohesin function would have made embryonic development inviable and thus could have never been observed on a patient. This fact makes CdLS causing variants affecting the cohesin complex potentially very informative about cohesin atomistic mechanisms, as these are necessarily affecting key processes of the cohesin dynamics while keeping enough functionality to let embryonic development result in a viable individual. From this perspective CdLS individuals live on a thin edge of biological viability and, if any rationally designed therapies are ever to be developed, these will likely arise from in-depth atomistic description of subtle mechanisms regulating the functions of the cohesin complex.

Apart from disease related variants, our models were used to analyze three yeast Smc1 residues that, when mutated, can bypass the need for Eco1 [139]. The first two (Smc1A-G1131 and Smc1A-D1163) were very close to the γ-phosphate group of the ATP molecule in AS2 and, therefore, mutations affecting this residues can be expected to alter ATP hydrolysis in this site. More interesting is the third residue, Smc1-L1129, as it is not located in such an evidently relevant position and, on top of that, the orthologous mutation in yeast Smc3 protein does not bypass the need for Eco1, a fact the authors who described these mutants found surprising [139]. As happened in the case of the human Smc3-Q1147E variant, the predicted allosteric coupling between AS1 and AS2 was key to propose a possible mechanistic explanation. Based on the behavior of the key ortholog human residues observed in our simulations, we proposed that Smc1-L1129 (yeast ortholog to human Smc1A-L1128) hydrophobic side chain would be stabilizing the hydrocarbon chain of Smc1-K1121 (yeast ortholog to human Smc1A-K1120), keeping it correctly oriented towards AS2. On the other hand, the mutation of the Smc3 ortholog would not exhibit this phenotypic effect as there is no evidence of it being involved in any kind of allosteric coupling between active sites.

By the same line of assigning degrees of relevance to certain residues, as was commented before, a group of eight positively charged residues (Smc1A-K59, Smc1A-R62, Smc1A-K149, Smc3-H55, Smc3-R61, Smc3-K105, Smc3-K106 and Smc3-K157) became exposed during MD simulations acquiring an arrangement compatible with DNA binding. Two of them (Smc3-K105 and Smc3-K106) were already related to this function in the literature [128,132,140,174] but it was the first time the other six were proposed to participate in this process. It is worth noting that the mutations of three of these six residues proposed to be involved in cohesin DNA-binding (Smc1A-K59, Smc1A-R62 and Smc3-H55Y) are related to either CdLS [122,142,143,169] or colorectal cancer [19,168].

Our framework also offered atomistic information about the key steps in the two ATPase reactions that take place in a cohesin complex. This information proves valuable for future rational drug development as it facilitates the design of drugs that would block the reaction at a certain points, being transition state analogs, powerful enzymatic inhibitors [181–183], a typical example. However, given the ubiquity of ATP hydrolysis in living organisms, finding a truly cohesin specific transition state analog seems unlikely. Still, finding molecules that can regulate cohesin activity may provide potential treatments to certain CdLS patients, novel anti-cancer therapies or new experimental tools. In hope of finding a cohesin specific drug that could effectively alter cohesin function we tried a different approach, focusing our attention in other protein features rather than exclusively

the active sites. The details of this work in progress approximation will be detailed in the third results section of this thesis: "Allosteric coupling inhibitor screening via molecular docking".

3.2.4 Materials and methods in cohesin modeling

3.2.4.1 Homology modeling

To build the homology model of the trimer formed by Smc1A-head, Smc3-head and Rad21-Cter the protein sequences used were SMC1A_HUMAN (UniProt code: Q14683, residues 1 to 175 and 1058 to 1223), Smc3-head: SMC3_HUMAN (UniProt code: Q9UQE7, residues 1 to 179 and 1045 to 1206), and Rad21-Cter: RAD21_HUMAN (UniProt code: O60216, residues 543 to 629) all of them available in the UniProtKB database. Three different scaffold structures were used in order to reproduce various features. Both active sites (AS1 and AS2), Rad21-Cter, the interface between Rad21-Cter and Smc1A-head and the interface between Smc1A-head and Smc3-head were modeled on the structure of a *Saccharomyces cerevisiae* homodimeric Smc1 (human Smc1A ortholog) ATPase head complex bound to the C-terminal domain of the yeast Scc1 (human Rad21 ortholog) (Protein Data Bank ID: 1W1W[133]). The Smc3-head structure was modeled on a *Saccharomyces cerevisiae* Smc3 (human Smc3 ortholog) monomer bound to the Scc1 N-terminal domain (PDB ID: 4UX3[131]). The positions of residues surrounding the active sites were refined using the 3D structure of a *Pyrococcus furiosus* Smc homodimer (PDB ID: 1XEX[178]) and the crystallographic water molecules present were added to the model. The ATPγS molecules used in 1W1W as substrate analogues to block the ATPase activity of the dimer[133] were replaced by either ATP or ADP. The model of the human variant Smc3-Q1147E was generated by replacing the apical amide group in the Smc3-Q1147 residue by a carboxylate group. The resulting model is compatible both with recently published structures of human cohesin head obtained by high-resolution electron microscopy[184] and a recent crystallographic structure of bacterial SMC[180].

3.2.4.2 Free MD simulations

AMBER14 molecular dynamics package[43] was used to perform all the free MD simulations as well as the thermalization, equilibration and stabilization phases that are described below. All the 3D models generated were solvated with periodic cuboid pre-equilibrated solvent boxes of TIP3P model water molecules[61] using the LEaP module of AMBER, setting 12 Å as the shortest distance between any atom present in the 3D model and the periodic box boundaries. H++ web server[51–53] was used to determine protonation states and Na^+ counterions were added to neutralize the charge of the systems. All the free MD simulations were performed in the NPT (constant pressure, constant temperature) ensemble, using the parm99 forcefield[45,46] and were run with CUDA parallelized binaries of the PMEMD program of the AMBER package. The SHAKE algorithm[185,186] was used, allowing a time step of 2 fs. After protonation, solvation and charge neutralization all the initial structures were relaxed over 15,000 steps of energy minimization with a cut-off of 12Å. Then, structures were thermalized during a 20 ps long heating phase in which temperature was raised from 0 to 300 K in 10 2 ps long temperature change steps, after each of which velocities were reassigned. During both minimization and heating

phases, Cα trace dihedrals were restrained with a 500 kcal mol^{-1} rad^{-2} force constant. After the heating phase trace dihedrals were relaxed during a 20 ps long stabilization phase. Once the structures were thermalized and had Cα trace restraints removed, 120 to 150 ns long productive MD simulations were computed for each of them. All the structures presented gaps in the Cα trace of Smc1A-head and Smc3-head towards the long coiled-coil region, which cannot be accurately modeled due to the lack of template structures. To prevent unwanted unfolding events in the short coiled-coil modeled regions these gaps were protected with distance restraints that were applied during all phases. To improve the sampling of catalytic configurations, catalytic water at both active sites were restrained to hydrolysis compatible geometries. This was achieved by restraining both the distance between the oxygen atom of the catalytic water and the phosphorous atom of the ATP γ-phosphate group below 3.5 Å and the angle formed by these two atoms and the oxygen atom of the ATP β-phosphate group between 160° and 180°. Both distance and angle restraints were defined using a flat-bottomed potential allowing free sampling of geometries among the defined boundaries. Catalytic water restraints were released prior to QM/MM MD simulations of the active sites.

3.2.4.3 QM/MM MD and QM/MM SMD simulations

The recently developed method Fireball/AMBER[87,88] was used to carry out the QM/MM MD simulations. Fireball/AMBER offers a combination of the AMBER molecular dynamics package[43] and Fireball, a local-orbital density-functional theory molecular dynamics technique[86]. The systems were divided into two regions, an MM region governed by AMBER MM calculations in a similar fashion as free MD simulations and a QM region described through Fireball QM calculations using a basis set of optimized numerical atomic-like orbitals (NAOs) with a single s orbital for H, sp3 orbitals for C, N and O, and sp3d5 orbitals for P, as used in previous works[88,112]. The interaction between the atoms belonging to the QM and MM regions was automatically handled by the Fireball/AMBER method. The time step used during QM/MM MD simulations was 0.5 fs and the initial structures used were taken from free MD simulations after these became stable.

3.2.4.4 3D free energy surfaces generation

3D free energy surfaces were obtained as described in previous literature[88,110–112] using biased QM/MM MD simulations in which the two reaction coordinates were fixed with restraints and forced to sample different regions of the conformational space over time like in SMD (QM/MM SMD). Sampling was divided into three phases. First phase consisted of moving both reaction coordinates to the lowest leftmost point region of the energy surface to be sampled (i.e. RC1: 1.5, RC2: 1.6) to obtain our initial sampling structure (Fig. 12). Second phase consisted of sampling along RC1 while keeping RC2 fixed (Fig. 12). In the third and last phase, structures created in the second phase were used generate QM/MM SMD trajectories sampling along RC2 while keeping RC1 fixed at uniformly distributed values. Following this method, 7.6 x 10^6 structures with their associated reaction coordinates and energy values were obtained. The QM energy values were distributed across a uniform 2D grid defined by the two reaction coordinates, creating groups of $\sim 1.5 \times 10^4$ different structures on average. To estimate a free energy surface the partition function was calculated for each group and the results were then

smoothed via a 3D LOESS local regression. The free energy surfaces obtained for each condition were analyzed using MEPSA[23] to obtain reaction paths and energy profiles. These paths were used to find the points of interest shown in (figures 51, 52 and 58). The path obtained in the free energy surface calculated in presence of Rad21-Cter was converted to MD restraints and used to generate a QM/MM SMD trajectory (appendix C).

3.2.4.5 2D free energy profiles generation

2D free energy profiles were obtained by sampling through a QM/MM SMD trajectory along the reaction coordinate sampling the uniformly distributed initial structures. These structures were relaxed along 5 ps by keeping the reaction coordinate fixed. Velocities were reassigned every 0.5 ps. The last 0.5 ps of each relaxation, which yielded 7.7 x 10^4 structures with their associated reaction coordinates and energy values, were used for 2D free energy profile generation. The QM energy values were distributed across a uniform 1D grid defined by the reaction coordinate, creating groups of $\sim 10^3$ different structures on average. To estimate a free energy surface the partition function was calculated for each group and the results were then smoothed via a 2D LOESS local regression.

3.2.4.6 Error analysis

Error analysis was performed by bootstrap resampling (100 replicates) on the data for both 3D free energy surfaces (Fig. 64) and 2D free energy profiles (Fig. 65). The standard deviation was found to be below 0.8 kcal mol^{-1} for 3D surfaces and 0.5 kcal mol^{-1} for 2D profiles in all the relevant positions.

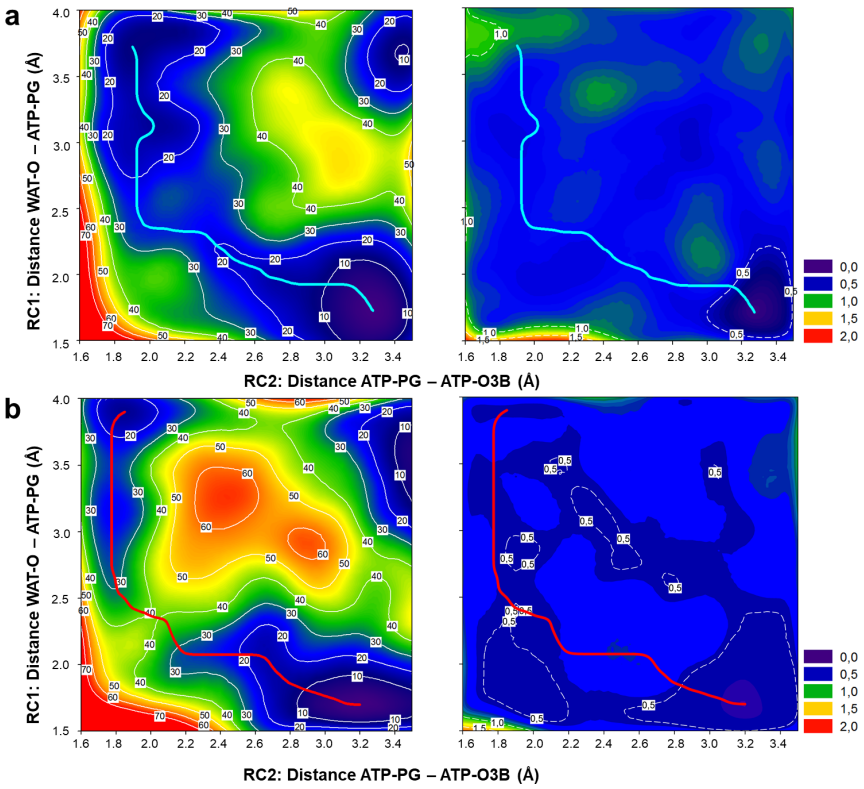


Figure 64: *3D free energy surfaces error estimation via bootstrapping (100 resampling replicates). Standard deviation values (right) of the free energy surface (left) for ATP hydrolysis at AS1 in the presence (**a**) and absence (**b**) of Rad21-Cter. Color scale for standard deviation values is included. Modified from Marcos-Alcalde et al. (2017)[24]. Caption was adapted from the same source.*

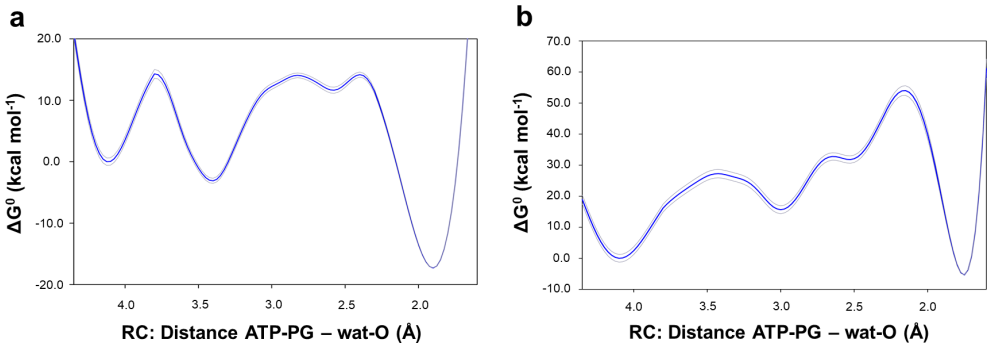


Figure 65: *2D free energy profiles error estimation via bootstrapping (100 resampling replicates). Mean values (blue lines) +/- standard deviation (gray lines) of the 2D free energy profile of ATP hydrolysis at AS2 in the AS1-ADP/AS2-ATP (**a**) and AS1-ATP/AS2-ATP (**b**) conditions. Modified from Marcos-Alcalde et al. (2017)[24]. Caption was adapted from the same source.*

3.2.4.7 Free energy difference calculations from SMD simulations

Free energy calculations from SMD simulations using Jarzynski's equality[84] were performed to compare the dissociation of Smc1A-head Smc3-head dimer either binding ATP (AS1-ATP/AS2-ATP condition) or ADP (AS1-ADP/AS2-ADP condition) in both active sites. Ten SMD trajectories, five for each condition, were generated starting from structures obtained from stable MD simulations. Before starting the SMD simulations, the periodic boundaries were expanded in the direction of the separation, preventing protein-protein collisions through them. Water box was enlarged accordingly by adding preequilibrated cuboid boxes of TIP3P water model molecules. After box enlargement, a similar heating protocol as the one described for free MD simulations was run, ensuring an adequate thermalization of the system. Along each SMD trajectory, the centers of mass of Smc1A-head and Smc3-head were forced to separate 32.5 Å at a constant velocity over 13 ns (2.5 Å ns^{-1}) with a spring constant of 5 kcal mol^{-1} $Å^{-2}$, which is in the range of conditions used in similar SMD studies[187,188]. To better establish quasi-equilibrium conditions in the initial and final states of the SMD trajectories, which is required to use Jarzynki's equality, separation distance was kept constant for 0.1 ns at the beginning and the end of each trajectory. Cα trace dihedrals were restrained with a 500 kcal mol^{-1} rad^{-2} force constant to prevent large rearrangements of the heads structure during SMD and the distance restraints protecting the gaps in the free MD simulations were kept. In order to reconstruct the force and work generated along each trajectory, the distance between the centers of mass was recorded on each calculation step. In figure 59 the initial distance between the centers of mass was taken as the origin (0.0 Å) of separation.

3.3 Allosteric coupling inhibitor screening via molecular docking

3.3.1 Introduction

Based on the simulations described in section 3.2 we proposed a novel molecular mechanism leading to allosteric coupling between cohesin active sites for the first time. Due to the atomistic information obtained, a previously unreported pocket located on the interface between Smc1A-head and Smc3-head could be observed (Fig. 66 a). Interestingly, when AS2 activation occurs via the entrance of K1120, this pocket significantly shrinks (Fig. 66 b). This observation led to the hypothesis that if such shrinking could be blocked, allosteric coupling mechanism would probably be impeded to some extent, thus likely resulting in a putative cohesin opening specific inhibitor. In order to evaluate possible drugs that could bind the open pocket a docking pipeline using smina[103] (a fork of Autodock Vina[104]) was developed, parallelizing the pipeline to make the most of the 48-core computer available. More details on the pipeline implementation can be found in subsection 1.2.4. The results presented in this section were obtained as a proof of concept and are part of a work-in-progress. They are shown as part of the future perspective derived from the work performed in this thesis.

3.3.2 Results

The library of molecules used in this first screening was the biogenic subset of the ZINC15 database[102]. Despite ZINC15 offers ready-to-dock formats for some entries, this feature is only available to less than a half of the compounds present in the database and is heavily biased toward small molecules, as their structures are much simpler to predict. Large and complex compounds are extensively present in the biogenic data set, so relying on ZINC15 ready-to-dock formats was not a realistic option. Our approach consisted on downloading the full 2D biogenic dataset, perform structural prediction with CORINA[189,190] and use the prepare_ligand.py script available in Autodock Tools[191] to generate flexible ligand structures compatible with smina. After 3D modeling, the library formed by 130374 molecules was screened with smina using two representative structures of the Smc1A-head Smc3-head interface pocket before (open; Fig. 66 a) and after (closed; Fig. 66 b) AS2 activation as receptors. For each receptor, docking screening was performed in three phases with increasingly detailed sampling combined with progressively restrictive binding energy thresholds. Under this scheme, on each phase the ligands that exhibit a binding energy higher than the defined cutoff are filtered out. The parameters used in each pass and the number of structures that successfully passed the binding energy threshold are shown in table 4.

The ligands that reached phase 3 in the open conformation where simultaneously clustered by structural similarity using a combination of the FP2, FP3 and MACCS fingerprints available in Open Babel[192] and sorted by lowest binding energy. Evaluating the binding energy and cluster size a list of five compounds was obtained, the most promising of which (SMC-INH-1 from now on; Fig. 66 c) showed the best simultaneous scoring on both criteria. This list of compounds was shared with the group of Dr. Pedro A. Lazo-Zbikowski Taracena at the CIC (Centro de Investigación del Cancer USAL-CSIC) with

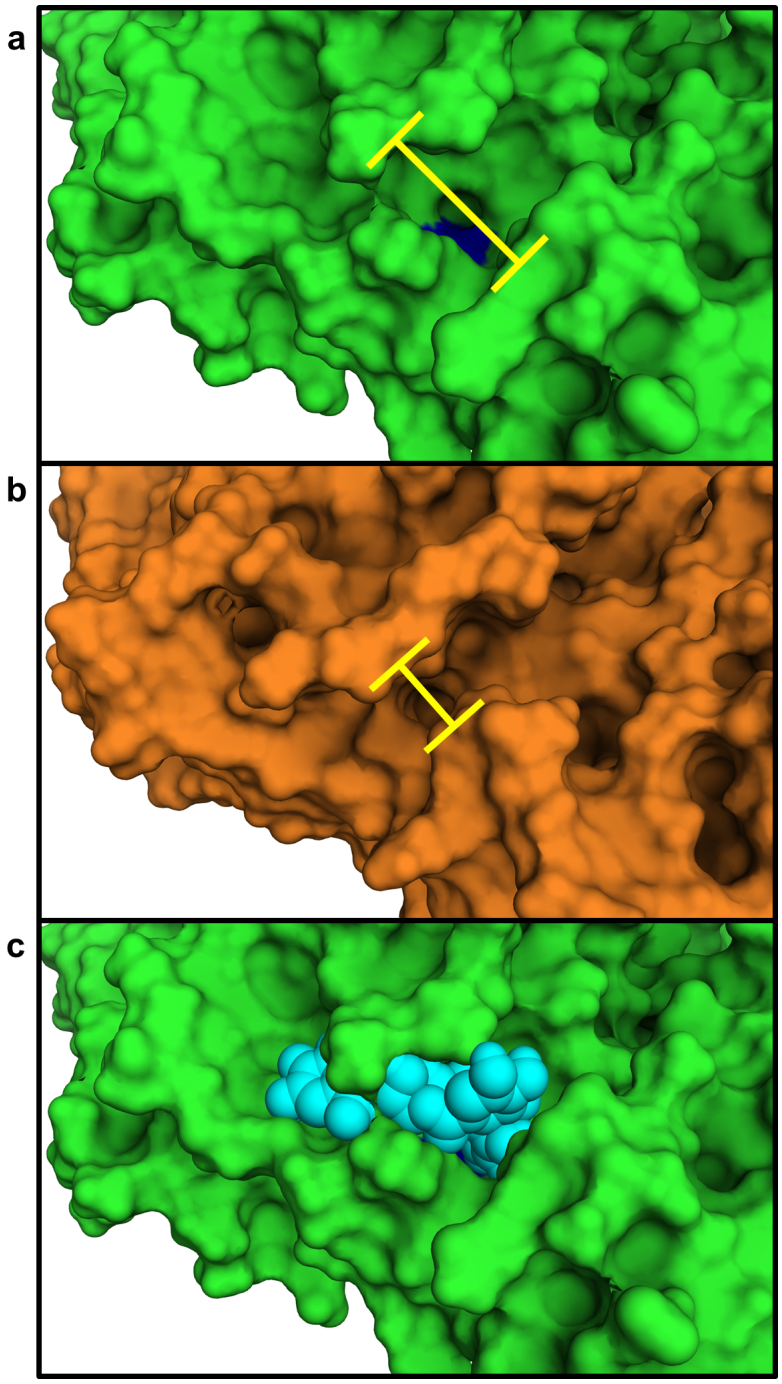


Figure 66: *Graphic depiction of the pocket that was used to perform allosteric coupling inhibitor screening via molecular docking. (**a**) Structure of the pocket in its open conformation before AS2 allosteric activation. Smc1A-K1120 location is indicated in blue.(**b**) Structure of the pocket in its closed conformation after AS2 allosteric activation. Yellow lines indicate the gap of the pocket that most notoriously shrinks during the conformational change. (**c**) Illustration of the most favorable predicted binding mode of SMC-INH-1 (cyan spheres) to the open pocket depicted in (**a**).*

whom a collaboration was started. They evaluated the effect of SMC-INH-1 over 293T cells (human embryonic kidney cell line carrying the SV40 T-antigen) via flow cytometry. Results confirmed that SMC-INH-1 induces cell cycle arrest in G2/M phase. This effect was evaluated both in asynchronic culture (Fig. 67) and in cells previously synchronized

	Phase 1	**Phase 2**	**Phase 3**
Exhaustiveness	2	6	8
Minimization steps	5	5	Auto scale
Binding energy cutoff	-6	-9	-10
Open conformation	83463	10945	3453
Closed conformation	75653	4567	854

Table 4: *Parameters used in docking protocol. The number of molecules that passed each phase for both targets (open conformation and closed conformation) is indicated.*

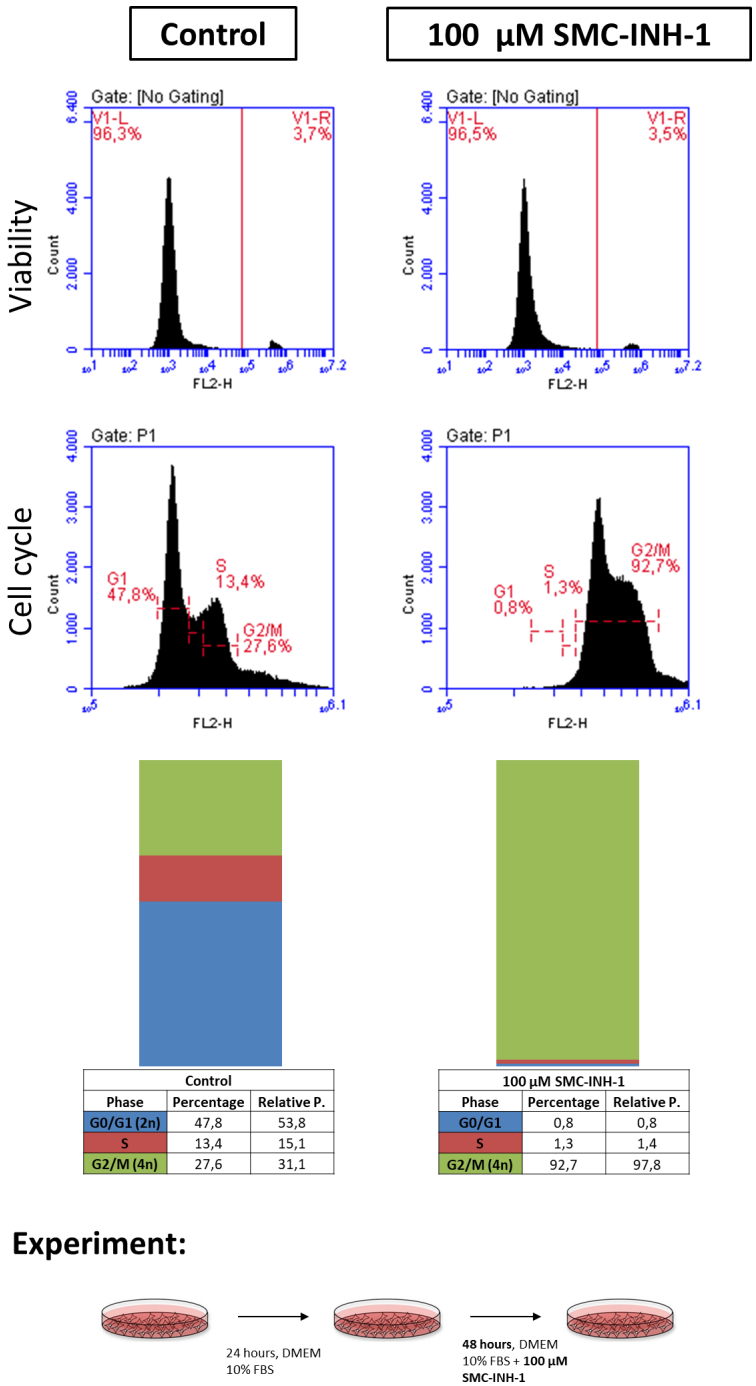
with nocodazole (Fig. 68). The proportion of cells in asynchronic culture at G2/M phase after SMC-INH-1 treatment rose from 31.1% to 97.8% (Fig. 67). In nocodazole pretreated culture, after nocodazole release, the proportion of cells in G2/M is 79.2% in presence of SMC-INH-1, in contrast to the 46% detected in absence of the compound (Fig. 68). These results clearly support that SMC-INH-1 induces G2/M arrest over 293T cells, which is compatible with the expected effect of inhibiting the allosteric coupling between cohesin head active sites.

3.3.3 Discussion

No drug-like cohesin specific inhibitor has been published to date. Our preliminary results suggest that SMC-INH-1 may be inducing G2/M arrest by hindering the allosteric coupling mechanism predicted between cohesin ATPase active sites. If further validated, this could make SMC-INH-1 a promising tool for both the study of cohesin dynamics as well as a novel drug for low-lethality G2/M arrest.

SMC-INH-1 was tested using 1 μM, 10 μM, 50 μM and 100 μM concentrations. The effect of 1 μM and 10 μM concentrations was unnoticeable (data not shown), 50 μM yielded subtle results (data not shown) and 100 μM had to be used to obtain clear G2/M arrest. Given the high dose required to obtain measurable effect, these results make SMC-INH-1 an unlikely therapeutic proposal on its own. However, if the predicted molecular mechanism of action is confirmed, the search of compounds with similar binding modes would prove a promising way to search for new molecular tools and therapeutic drugs[193].

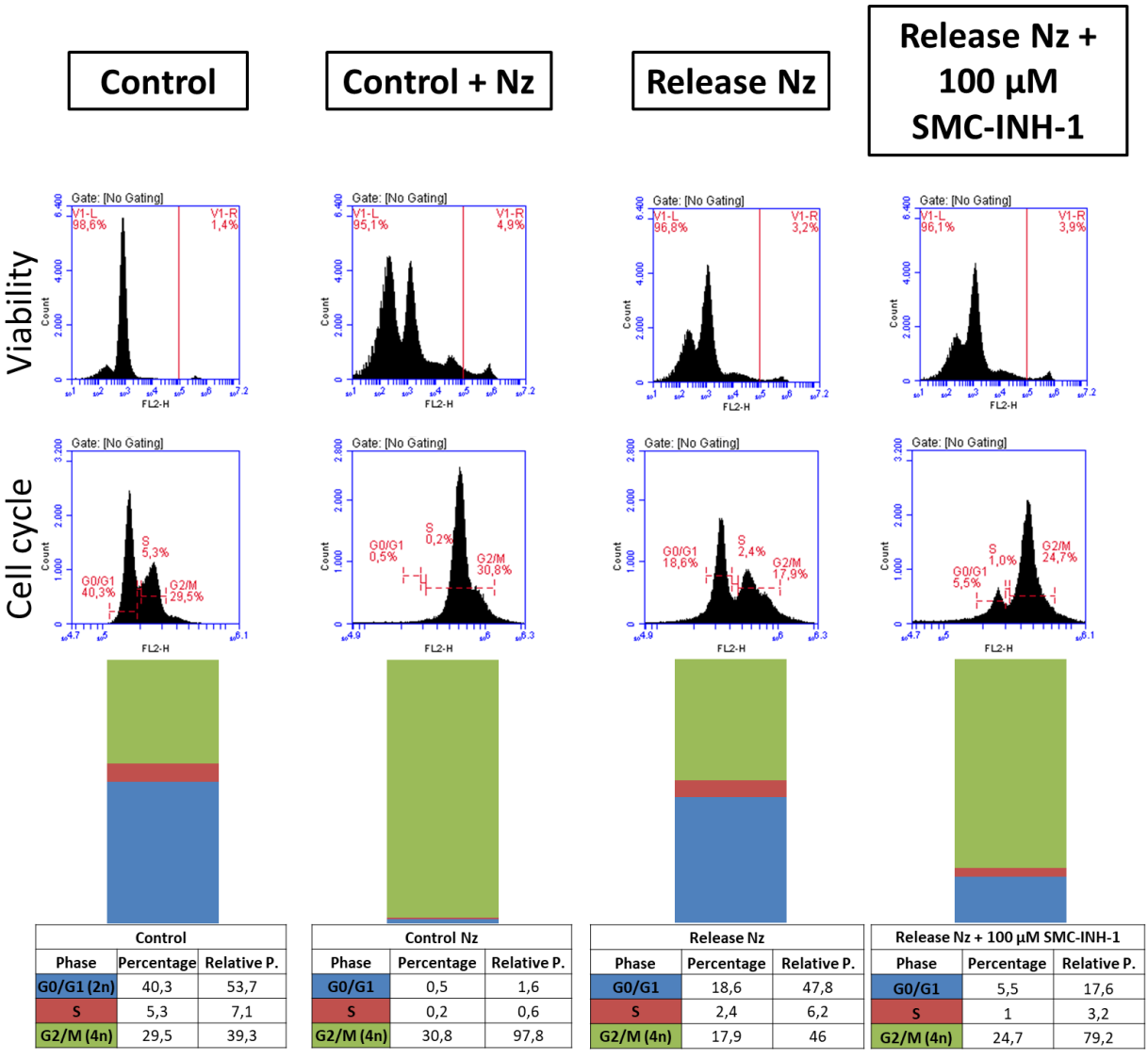
The data presented here is part of a work-in-progress that will require further development of both the molecular docking protocol and the molecular characterization of the measured effect. On one hand, the current docking pipeline will be optimized and extended to handle larger datasets. On the other, the molecular mechanism governing G2/M arrest by SMC-INH-1 will have to be thoroughly characterized in order to confirm that the predicted binding mode is the actual cause of the observed G2/M arrest. If molecular confirmation of this mechanism is obtained, systematic experimental screening of compounds with similar predicted binding modes would commence in hope of finding candidates with better dose/effect ratio that could possibly lead to better G2/M arrest inducing molecular tools and, most of all, actual therapeutic candidates.



Control		
Phase	Percentage	Relative P.
G0/G1 (2n)	47,8	53,8
S	13,4	15,1
G2/M (4n)	27,6	31,1

100 µM SMC-INH-1		
Phase	Percentage	Relative P.
G0/G1	0,8	0,8
S	1,3	1,4
G2/M (4n)	92,7	97,8

Figure 67: *Cell cycle arrest in asynchronic 293T cells treated with 100 µM SMC-INH-1 for 48 hours. FACS output is shown in the top pannels. Total and relative percentages of cells in each cell cycle phase for each condition are shown in tables. Relative percentages are also represented in column charts following the color scheme introduced in the tables. At the bottom, a graphical depiction of the experimental protocol is presented.*



Control		
Phase	Percentage	Relative P.
G0/G1 (2n)	40,3	53,7
S	5,3	7,1
G2/M (4n)	29,5	39,3

Control Nz		
Phase	Percentage	Relative P.
G0/G1	0,5	1,6
S	0,2	0,6
G2/M (4n)	30,8	97,8

Release Nz		
Phase	Percentage	Relative P.
G0/G1	18,6	47,8
S	2,4	6,2
G2/M (4n)	17,9	46

Release Nz + 100 µM SMC-INH-1		
Phase	Percentage	Relative P.
G0/G1	5,5	17,6
S	1	3,2
G2/M (4n)	24,7	79,2

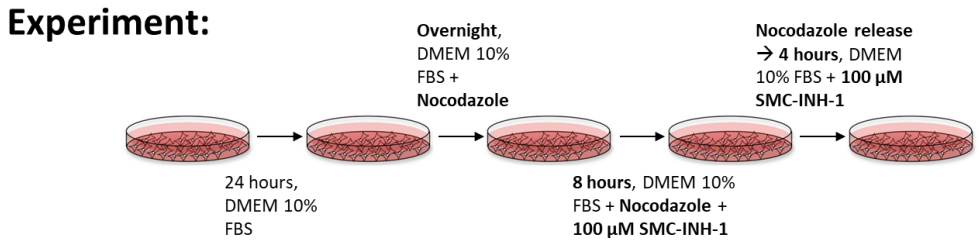


Figure 68: *Cell cycle arrest in 293T cells previously synchronized with nocodazole (Nz) and then treated with 100 µM SMC-INH-1 for 4 hours. FACS output is shown in the top pannels. Total and relative percentages of cells in each cell cycle phase for each condition are shown in tables. Relative percentages are also represented in column charts following the color scheme introduced in the tables. At the bottom, a graphical depiction of the experimental protocol is presented.*

4 | Discussion & Conclusions

The description of molecular interactions and chemical reactions in atomistic terms provides a powerful tool to both explain and predict phenotype-inducing mutations as well as perform rational drug design or screening.

In this thesis, a combination of computational approaches (homology modeling, MD, QM/MM MD, SMD and molecular docking) were used to gain atomistic insight into the dynamics of the ATPase head heterodimer of human cohesin complex, a key actor in processes of sister chromatid cohesion, DNA repair, chromatin organization and transcription regulation. Variants on members of this complex are connected to a series of rare diseases referred to as cohesinopathies as well as several types of cancer.

Additionally, a computational tool to perform user-friendly analysis of QM/MM MD 3D free energy surfaces (a central type of data generated through the pipeline presented in this thesis) was developed and published.

Lastly, the atomistic description of cohesin ATPase head dynamics allowed us to perform *in silico* screening of potential cohesin inhibitors against a previously unreported molecular target, yielding promising preliminary results.

To provide a structured list of the conclusions derived from this thesis, these will be presented bellow following the organization of the results chapter (3).

Regarding the results presented in section 3.1 "MEPSA: minimum energy pathway analysis for energy landscapes", the major conclusion that can be presented is:

- MEPSA offers a user-friendly tool to analyze 3D energy surfaces that significantly facilitates both data interpretation and figure generation.

The results exposed in section 3.2 "Two-step ATP-driven opening of cohesin head" support a series of conclusions that can be sub-divided into two groups: those describing general mechanisms of the cohesin dynamics and those pointing out features of particular residues.

- Conclusions regarding general mechanism of the cohesin dynamics:
  - Rad21-Cter binding to Smc1A-head facilitates ATP hydrolysis in active site 1 by inducing a rearrangement in this site that both facilitates the entrance of the catalytic water molecule and stabilizes the transition state, significantly reducing the free energy barrier associated with the reaction.
  - Cohesin exhibits allosteric coupling between active site 1 and 2 as the presence of ADP in active site 1 induces an internal rearrangement in the protein complex that facilitates ATP hydrolysis in active site 2.

  - ATP hydrolysis in both active sites of the cohesin head complex strongly debilitates its stability, thus leading to ring opening.

- Conclusions regarding features of particular residues in the context of cohesin dynamics:
  - The presence of ADP in active site 1 induces the entrance of Smc1A-K1120 in active site 2 which both facilitates the entrance of the catalytic water molecule and stabilizes the transition state, proving Smc1A-K1120 to be a major actor in the allosteric coupling between active site 1 and 2.
  - CdLS causing Smc3-Q1147E variant partially interferes with the allosteric coupling between active site 1 and 2, producing an impaired, yet not completely disrupted, behavior of the cohesin complex compatible with the patient phenotype.
  - The Eco1 bypass causing yeast mutation Smc1-L1129V interferes with the allosteric coupling between active site 1 and 2, which both explains the Eco1 bypass and the neutral character of the equivalent mutation in Smc3.
  - Residues Smc1A-K59, Smc1A-R62, Smc1A-K149, Smc3-H55, Smc3-R61 and Smc3-K157 are proposed to participate in the DNA binding process of the cohesin complex.

The work-in-progress results present in section 3.3 "Allosteric coupling inhibitor screening via molecular docking" support the following conclusions:

- SMC-INH-1, a compound computationally predicted to be an allosteric coupling inhibitor induces G2/M cell cycle arrest in human 293T cells.

5 | Discusión y Conclusiones

La descripción de interacciones moleculares y reacciones químicas a escala atómica ofrece una potente herramienta en la explicación y predicción de fenotipos inducidos por mutaciones así como en la búsqueda y el diseño racional de fármacos.

En esta tesis se ha empleado una combinación de métodos computacionales (modelado por homología, diferentes técnicas de dinámica molecular y docking molecular) para estudiar la dinámica asociada a la actividad ATPasa de las cabezas del complejo heterodimérico que forma la cohesina humana, una pieza fundamental en procesos de cohesion de cromátidas hermanas, reparación de daño en el ADN, organización de la cromatina y regulación transcripcional. Variantes en miembros de este complejo se han vinculado a un conjunto de enfermedades raras denominadas cohesinopatías al igual que a diferentes tipos de cáncer.

Además se ha desarrollado una herramienta informática que permite analizar superficies de energía libre tridimensionales (un tipo de dato central en esta tesis) de forma sencilla y accesible.

Por último, la descripción a escala atómica de la dinámica asociada a la actividad ATPasa de las cabezas de la cohesina humana ha hecho posible la búsqueda computacional de inhibidores potenciales específicos frente a esta nueva diana molecular. Los resultados preliminares derivados de esta aproximación son prometedores.

Con el fin de ofrecer una lista estructurada de las conclusiones derivadas de esta tesis, éstas se mostrarán a continuación ordenadas de acuerdo a la estructura seguida a lo largo del capítulo de resultados (3):

En referencia a los resultados presentados en la sección 3.1 "MEPSA: minimum energy pathway analysis for energy landscapes", la principal conclusión que se puede extraer es:

- MEPSA ofrece una herramienta para análisis de superficies tridimensionales de energía libre accesible, que facilita tanto el análisis de este tipo de datos como la posterior generación de figuras.

Los resultados expuestos en la sección 3.2 "Two-step ATP-driven opening of cohesin head" apoyan una serie de conclusiones que pueden subdividirse en dos grupos: aquellas que describen mecanismos generales de la dinámica de cohesinas y aquellas que permiten detallar características de residuos individuales.

- Conclusiones referentes a mecanismos generales de la dinámica de cohesinas:
  - La unión del dominio C-terminal de Rad21 a la cabeza de Smc1A facilita la

hidrólisis de ATP en el centro activo 1 al inducir una reorganización estructural en dicho sitio que favorece la entrada de la molécula de agua catalítica y la estabilización del estado de transición, bajando significativamente la barrera energética de la reacción.

  - Existe un acoplamiento alostérico entre los centros activos 1 y 2 del complejo cohesina. La presencia de ADP en el centro activo 1 induce una reorganización interna del complejo que facilita la hidrólisis de ATP en el centro activo 2.
  - La hidrólisis de ATP en los dos centros activos del complejo cohesina desestabiliza la interacción de las cabezas y, por tanto, favorece la apertura del anillo.

- Conclusiones referentes a residuos concretos en el contexto de la dinámica de cohesinas:
  - La presencia de ATP en el centro activo 1 induce la entrada del residuo Smc1A-K1120 en el centro activo 2, facilitando la entrada del agua catalítica y la estabilización del estado de transición en este último. Por tanto, Smc1A-K1120 es un actor fundamental en el acoplamiento alostérico entre los centros activos 1 y 2.
  - La variante causante de Síndrome de Cornelia de Lange Smc3-Q1147E interfiere parcialmente con el acoplamiento alostérico entre los centros activos 1 y 2, dando lugar a un funcionamiento defectuoso de la enzima sin llegar a anularlo por completo, lo cual es compatible con el fenotipo observado en el paciente.
  - La mutación Smc1-L1129V, que en levadura genera un fenotipo resistente a la pérdida de Eco1, interfiere con el acoplamiento alostérico entre los centros activos 1 y 2, lo cual explicaría tanto el fenotipo de resistencia a la pérdida de Eco1 observado en este mutante como el fenotipo neutro que presenta la mutación equivalente en Smc3.
  - Se propone que los residuos Smc1A-K59, Smc1A-R62, Smc1A-K149, Smc3-H55, Smc3-R61 y Smc3-K157 podrían participar en el proceso de unión de DNA del complejo cohesina.

Los resultados preliminares mostrados en la sección 3.3 "Allosteric coupling inhibitor screening via molecular docking" permiten sostener las siguientes conclusiones:

- SMC-INH-1, un compuesto que se ha predicho computacionalmente como un inhibidor del acoplamiento alostérico, induce arresto del ciclo celular en G2/M en células 293T humanas.

6 | Future perspectives

The results presented in this thesis open a wide range of future perspectives from which further develop this work.

Regarding MEPSA we are particularly excited about generalizing its core functionalities to work with n-dimensional energy surfaces. Computational methods such as metadynamics, which are able to calculate free energy surfaces over an arbitrary number of dimensions with affordable computational costs, are becoming increasingly popular. MEPSA n-dimensional generalization would dramatically simplify the analysis of data that otherwise may result quite challenging to analyze due to its sheer complexity.

In relation to cohesin mechanics, we would like to characterize the driving force of Smc1A-K1120 mediated allosteric coupling in detail. We are also interested in thoroughly evaluating the effect the interaction with double helix DNA may have both on the geometry of the active sites and on the allosteric coupling mechanism. If such DNA-protein interaction can be successfully modeled, the effect of lysine acetylations could be thoroughly analyzed. Also, this model could prove useful in case any relevant directional bias in the DNA sliding motion over the head complex is observed, as that could provide an atomistic mechanism for loop extrusion. In addition, although it lies beyond our current reach, experimental evaluation of the allosteric coupling mechanism would certainly be desirable. Lastly, after all the accumulated experience simulating the cohesin head heterodimer, it seems quite convenient to reproduce the same approach on human condensin, being the evaluation of the allosteric coupling mechanism one of the most intriguing questions for us.

With respect to the presented drug discovery pipeline, it has to be further optimized and other ZINC15 subsets have to be evaluated. Apart from computational considerations, the experimental validation of the proposed inhibitory mechanism for SMC-INH-1 is especially relevant yet certainly challenging. In case it is experimentally confirmed, not only it would confirm the discovery of the first cohesin ATPase inhibitor to date, but would also make the search for molecules with similar binding modes therapeutically promising to some extent. Finally, a more immediate approach regarding these results that we are already taking is to make use of the developed pipeline with other proteins that may have biological and/or medical interest, trying to identify other molecules with potential applications. In this regard, we have currently identified two potential inhibitors of the GTPase activity of the bacterial FtsZ that exhibit *in-vitro* inhibitory dose/effect ratios similar to that of the antibiotic kanamycin.

Bibliography

[1] B. N. Kholodenko. Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol*, 7(3):165–76, March 2006.

[2] A. Warshel. Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu Rev Biophys Biomol Struct*, 32:425–43, 2003.

[3] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*, 44(D1):D279–85, January 2016.

[4] K. P. Hopfner, A. Karcher, D. S. Shin, L. Craig, L. M. Arthur, J. P. Carney, and J. A. Tainer. Structural biology of Rad50 ATPase: ATP-driven conformational control in DNA double-strand break repair and the ABC-ATPase superfamily. *Cell*, 101(7):789–800, June 2000.

[5] J. Lowe, S. C. Cordell, and F. van den Ent. Crystal structure of the SMC head domain: an ABC ATPase with 900 residues antiparallel coiled-coil inserted. *J Mol Biol*, 306(1):25–35, February 2001.

[6] K. Jeppsson, T. Kanno, K. Shirahige, and C. Sjogren. The maintenance of chromosome structure: positioning and functioning of SMC complexes. *Nat Rev Mol Cell Biol*, 15(9):601–14, September 2014.

[7] T. Gligoris and J. Lowe. Structural Insights into Ring Formation of Cohesin and Related Smc Complexes. *Trends Cell Biol*, 26(9):680–93, September 2016.

[8] J. Pie, B. Puisac, M. Hernandez-Marcos, M. E. Teresa-Rodrigo, M. Gil-Rodriguez, C. Baquero-Montoya, M. Ramos-Caceres, M. Bernal, A. Ayerza-Casas, I. Bueno, P. Gomez-Puertas, and F. J. Ramos. Special cases in Cornelia de Lange syndrome: The Spanish experience. *Am J Med Genet C Semin Med Genet*, 172(2):198–205, June 2016.

[9] M. C. Gil-Rodriguez, M. A. Deardorff, M. Ansari, C. A. Tan, I. Parenti, C. Baquero-Montoya, L. B. Ousager, B. Puisac, M. Hernandez-Marcos, M. E. Teresa-Rodrigo, I. Marcos-Alcalde, J. J. Wesselink, S. Lusa-Bernal, E. K. Bijlsma, D. Braunholz, I. Bueno-Martinez, D. Clark, N. S. Cooper, C. J. Curry, R. Fisher, A. Fryer, J. Ganesh, C. Gervasini, G. Gillessen-Kaesbach, Y. Guo, H. Hakonarson, R. J. Hopkin, M. Kaur, B. J. Keating, M. Kibaek, E. Kinning, T. Kleefstra, A. D. Kline, E. Kuchinskaya, L. Larizza, Y. R. Li, X. Liu, M. Mariani, J. D. Picker, A. Pie, J. Pozojevic, E. Queralt, J. Richer, E. Roeder, A. Sinha, R. H. Scott, J. So, K. A. Wusik, L. Wilson, J. Zhang, P. Gomez-Puertas, C. H. Casale, L. Strom, A. Selicorni, F. J.

Ramos, L. G. Jackson, I. D. Krantz, S. Das, R. C. Hennekam, F. J. Kaiser, D. R. FitzPatrick, and J. Pie. De novo heterozygous mutations in SMC3 cause a range of Cornelia de Lange syndrome-overlapping phenotypes. *Hum Mutat*, 36(4):454–62, April 2015.

[10] E. Watrin, F. J. Kaiser, and K. S. Wendt. Gene regulation and chromatin organization: relevance of cohesin mutations to human disease. *Curr Opin Genet Dev*, 37:59–66, April 2016.

[11] F. J. Ramos, B. Puisac, C. Baquero-Montoya, M. C. Gil-Rodriguez, I. Bueno, M. A. Deardorff, R. C. Hennekam, F. J. Kaiser, I. D. Krantz, A. Musio, A. Selicorni, D. R. FitzPatrick, and J. Pie. Clinical utility gene card for: Cornelia de Lange syndrome. *Eur J Hum Genet*, 23(10), October 2015.

[12] P. Chetaille, C. Preuss, S. Burkhard, J. M. Cote, C. Houde, J. Castilloux, J. Piche, N. Gosset, S. Leclerc, F. Wunnemann, M. Thibeault, C. Gagnon, A. Galli, E. Tuck, G. R. Hickson, N. El Amine, I. Boufaied, E. Lemyre, P. de Santa Barbara, S. Faure, A. Jonzon, M. Cameron, H. C. Dietz, E. Gallo-McFarlane, D. W. Benson, C. Moreau, D. Labuda, S. H. Zhan, Y. Shen, M. Jomphe, S. J. Jones, J. Bakkers, and G. Andelfinger. Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm. *Nat Genet*, 46(11):1245–9, November 2014.

[13] K. Izumi, R. Nakato, Z. Zhang, A. C. Edmondson, S. Noon, M. C. Dulik, R. Rajagopalan, C. P. Venditti, K. Gripp, J. Samanich, E. H. Zackai, M. A. Deardorff, D. Clark, J. L. Allen, D. Dorsett, Z. Misulovin, M. Komata, M. Bando, M. Kaur, Y. Katou, K. Shirahige, and I. D. Krantz. Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin. *Nat Genet*, 47(4):338–44, April 2015.

[14] L. Mannini, S. Menga, and A. Musio. The expanding universe of cohesin functions: a new genome stability caretaker involved in human disease and cancer. *Hum Mutat*, 31(6):623–30, June 2010.

[15] X. W. Pan, S. S. Gan, J. Q. Ye, Y. H. Fan, Upsilon Hong, C. M. Chu, Y. Gao, L. Li, X. Liu, L. Chen, Y. Huang, H. Xu, J. Z. Ren, L. Yin, F. J. Qu, H. Huang, X. G. Cui, and D. F. Xu. SMC1a promotes growth and migration of prostate cancer in vitro and in vivo. *Int J Oncol*, 49(5):1963–1972, November 2016.

[16] D. A. Solomon, J. S. Kim, and T. Waldman. Cohesin gene mutations in tumorigenesis: from discovery to clinical significance. *BMB Rep*, 47(6):299–310, June 2014.

[17] M. Krauthammer, Y. Kong, B. H. Ha, P. Evans, A. Bacchiocchi, J. P. McCusker, E. Cheng, M. J. Davis, G. Goh, M. Choi, S. Ariyan, D. Narayan, K. Dutton-Regester, A. Capatana, E. C. Holman, M. Bosenberg, M. Sznol, H. M. Kluger, D. E. Brash, D. F. Stern, M. A. Materin, R. S. Lo, S. Mane, S. Ma, K. K. Kidd, N. K. Hayward, R. P. Lifton, J. Schlessinger, T. J. Boggon, and R. Halaban. Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet*, 44(9):1006–14, September 2012.

[18] K. H. Metzeler, T. Herold, M. Rothenberg-Thurley, S. Amler, M. C. Sauerland, D. Gorlich, S. Schneider, N. P. Konstandin, A. Dufour, K. Braundl, B. Ksienzyk, E. Zellmeier, L. Hartmann, P. A. Greif, M. Fiegl, M. Subklewe, S. K. Bohlander, U. Krug, A. Faldum, W. E. Berdel, B. Wormann, T. Buchner, W. Hiddemann,

J. Braess, and K. Spiekermann. Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood*, 128(5):686–98, August 2016.

[19] S. Seshagiri, E. W. Stawiski, S. Durinck, Z. Modrusan, E. E. Storm, C. B. Conboy, S. Chaudhuri, Y. Guan, V. Janakiraman, B. S. Jaiswal, J. Guillory, C. Ha, G. J. Dijkgraaf, J. Stinson, F. Gnad, M. A. Huntley, J. D. Degenhardt, P. M. Haverty, R. Bourgon, W. Wang, H. Koeppen, R. Gentleman, T. K. Starr, Z. Zhang, D. A. Largaespada, T. D. Wu, and F. J. de Sauvage. Recurrent R-spondin fusions in colon cancer. *Nature*, 488(7413):660–4, August 2012.

[20] V. K. Hill, J. S. Kim, and T. Waldman. Cohesin mutations in human cancer. *Biochim Biophys Acta*, 1866(1):1–11, August 2016.

[21] M. S. Williams and T. C. P. Somervaille. Leukemogenic Activity of Cohesin Rings True. *Cell Stem Cell*, 17(6):642–644, December 2015.

[22] M. De Koninck and A. Losada. Cohesin Mutations in Cancer. *Cold Spring Harb Perspect Med*, 6(12), December 2016.

[23] I. Marcos-Alcalde, J. Setoain, J. I. Mendieta-Moreno, J. Mendieta, and P. Gomez-Puertas. MEPSA: minimum energy pathway analysis for energy landscapes. *Bioinformatics*, 31(23):3853–5, December 2015.

[24] I. Marcos-Alcalde, J. I. Mendieta-Moreno, B. Puisac, M. C. Gil-Rodriguez, M. Hernandez-Marcos, D. Soler-Polo, F. J. Ramos, J. Ortega, J. Pie, J. Mendieta, and P. Gomez-Puertas. Two-step ATP-driven opening of cohesin head. *Sci Rep*, 7(1):3266, June 2017.

[25] O. U. Nalbantoglu. Dynamic programming. *Methods Mol Biol*, 1079:3–27, 2014.

[26] O. Gotoh. Heuristic alignment methods. *Methods Mol Biol*, 1079:29–43, 2014.

[27] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson, and D. G. Higgins. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, 7:539, October 2011.

[28] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez. A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res*, 38(Web Server issue):W695–9, July 2010.

[29] H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez. Analysis Tool Web Services from the EMBL-EBI. *Nucleic Acids Res*, 41(Web Server issue):W597–600, July 2013.

[30] J. Soding. Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 21(7):951–60, April 2005.

[31] G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D. G. Higgins. Sequence embedding for fast construction of guide trees for multiple sequence alignment. *Algorithms Mol Biol*, 5:21, May 2010.

[32] F. Sievers and D. G. Higgins. Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol*, 1079:105–16, 2014.

[33] N. V. Grishin. Fold change in evolution of protein structures. *J Struct Biol*, 134(2-3):167–85, May 2001.

[34] C. Venclovas. Methods for sequence-structure alignment. *Methods Mol Biol*, 857:55–82, 2012.

[35] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, October 1990.

[36] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 46(D1):D8–D13, January 2018.

[37] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, January 2000.

[38] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. Gallo Cassarino, M. Bertoni, L. Bordoli, and T. Schwede. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*, 42(Web Server issue):W252–8, July 2014.

[39] M. Remmert, A. Biegert, A. Hauser, and J. Soding. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods*, 9(2):173–5, December 2011.

[40] K. Arnold, L. Bordoli, J. Kopp, and T. Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, January 2006.

[41] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with NAMD. *J Comput Chem*, 26(16):1781–802, December 2005.

[42] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J Chem Theory Comput*, 4(3):435–47, March 2008.

[43] D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, CheathamI-IIT.E., T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu, and P.A. Kollman. Amber 14, 2014.

[44] B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *J Comput Chem*, 30(10):1545–614, July 2009.

[45] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling.

Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–25, November 2006.

[46] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99sb. *J Chem Theory Comput*, 11(8):3696–713, August 2015.

[47] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, 81(8):3684–3690, October 1984.

[48] T. J. Dolinsky, J. E. Nielsen, J. A. McCammon, and N. A. Baker. PDB2pqr: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res*, 32(Web Server issue):W665–7, July 2004.

[49] E. G. Alexov and M. R. Gunner. Incorporating protein conformational flexibility into the calculation of pH-dependent protein properties. *Biophys J*, 72(5):2075–93, May 1997.

[50] R. E. Georgescu, E. G. Alexov, and M. R. Gunner. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. *Biophys J*, 83(4):1731–48, October 2002.

[51] J. C. Gordon, J. B. Myers, T. Folta, V. Shoja, L. S. Heath, and A. Onufriev. H++: a server for estimating pKas and adding missing hydrogens to macromolecules. *Nucleic Acids Res*, 33(Web Server issue):W368–71, July 2005.

[52] J. Myers, G. Grothaus, S. Narayanan, and A. Onufriev. A simple clustering algorithm can be accurate enough for use in calculations of pKs in macromolecules. *Proteins*, 63(4):928–38, June 2006.

[53] R. Anandakrishnan, B. Aguilar, and A. V. Onufriev. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res*, 40(Web Server issue):W537–41, July 2012.

[54] D. M. Zuckerman. Equilibrium sampling in biomolecular simulations. *Annu Rev Biophys*, 40:41–62, 2011.

[55] L. C.T. Pierce, R. Salomon-Ferrer, C. de Olivera, C. A. F., J. A. McCammon, and R. C. Walker. Routine Access to Millisecond Time Scale Events with Accelerated Molecular Dynamics. *Journal of Chemical Theory and Computation*, 8(9):2997–3002, September 2012.

[56] V. Tsui and D. A. Case. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *Journal of the American Chemical Society*, 122(11):2489–2498, March 2000.

[57] M. Feig. Kinetics from Implicit Solvent Simulations of Biomolecules as a Function of Viscosity. *J Chem Theory Comput*, 3(5):1734–48, September 2007.

[58] R. E. Amaro, X. Cheng, I. Ivanov, D. Xu, and J. A. McCammon. Characterizing loop dynamics and ligand recognition in human- and avian-type influenza neuraminidases via generalized born molecular dynamics and end-point free energy calculations. *J Am Chem Soc*, 131(13):4702–9, April 2009.

[59] R. E. Amaro, R. V. Swift, L. Votapka, W. W. Li, R. C. Walker, and R. M. Bush. Mechanism of 150-cavity formation in influenza neuraminidase. *Nat Commun*, 2:388, July 2011.

[60] R. Anandakrishnan, A. Drozdetski, R. C. Walker, and A. V. Onufriev. Speed of conformational change: comparing explicit and implicit solvent molecular dynamics simulations. *Biophys J*, 108(5):1153–64, March 2015.

[61] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, July 1983.

[62] D. J. Price and C. L. Brooks. A modified TIP3p water potential for simulation with Ewald summation. *The Journal of Chemical Physics*, 121(20):10096–10103, November 2004.

[63] W. L. Jorgensen and J. D. Madura. Temperature and size dependence for Monte Carlo simulations of TIP4p water. *Molecular Physics*, 56(6):1381–1392, December 1985.

[64] H. W. Horn, W. C. Swope, J. W. Pitera, J. D. Madura, T. J. Dick, G. L. Hura, and T. Head-Gordon. Development of an improved four-site water model for biomolecular simulations: TIP4p-Ew. *J Chem Phys*, 120(20):9665–78, May 2004.

[65] H. W. Horn, W. C. Swope, and J. W. Pitera. Characterization of the TIP4p-Ew water model: vapor pressure and boiling point. *J Chem Phys*, 123(19):194504, November 2005.

[66] M. W. Mahoney and W. L. Jorgensen. A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions. *The Journal of Chemical Physics*, 112(20):8910–8922, May 2000.

[67] S. Izadi, R. Anandakrishnan, and A. V. Onufriev. Building Water Models: A Different Approach. *The Journal of Physical Chemistry Letters*, 5(21):3863–3871, November 2014.

[68] J. W. Caldwell and P. A. Kollman. Structure and Properties of Neat Liquids Using Nonadditive Molecular Dynamics: Water, Methanol, and N-Methylacetamide. *The Journal of Physical Chemistry*, 99(16):6208–6219, April 1995.

[69] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. The missing term in effective pair potentials. *The Journal of Physical Chemistry*, 91(24):6269–6271, November 1987.

[70] K. Takemura and A. Kitao. Water model tuning for improved reproduction of rotational diffusion and NMR spectral density. *J Phys Chem B*, 116(22):6279–87, June 2012.

[71] L. P. Wang, T. J. Martinez, and V. S. Pande. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *J Phys Chem Lett*, 5(11):1885–91, June 2014.

[72] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E.

Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102(18):3586–616, April 1998.

[73] R. B. Best, X. Zhu, J. Shim, P. E. Lopes, J. Mittal, M. Feig, and A. D. Mackerell, Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. *J Chem Theory Comput*, 8(9):3257–3273, September 2012.

[74] J. S. Hub, B. L. de Groot, H. Grubmuller, and G. Groenhof. Quantifying Artifacts in Ewald Simulations of Inhomogeneous Systems with a Net Charge. *J Chem Theory Comput*, 10(1):381–90, January 2014.

[75] K. Lindorff-Larsen, P. Maragakis, S. Piana, and D. E. Shaw. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *J Phys Chem B*, 120(33):8313–20, August 2016.

[76] T. D. Pollard and G. G. Borisy. Cellular motility driven by assembly and disassembly of actin filaments. *Cell*, 112(4):453–65, February 2003.

[77] H. S. Zaher and R. Green. Fidelity at the molecular level: lessons from protein synthesis. *Cell*, 136(4):746–62, February 2009.

[78] T. Pape, W. Wintermeyer, and M. V. Rodnina. Complete kinetic mechanism of elongation factor Tu-dependent binding of aminoacyl-tRNA to the A site of the E. coli ribosome. *EMBO J*, 17(24):7490–7, December 1998.

[79] T. M. Yi, H. Kitano, and M. I. Simon. A quantitative characterization of the yeast heterotrimeric G protein cycle. *Proc Natl Acad Sci U S A*, 100(19):10764–9, September 2003.

[80] M. C. Leake, J. H. Chandler, G. H. Wadhams, F. Bai, R. M. Berry, and J. P. Armitage. Stoichiometry and turnover in single, functioning membrane protein complexes. *Nature*, 443(7109):355–8, September 2006.

[81] M. Dunaway, J. S. Olson, J. M. Rosenberg, O. B. Kallai, R. E. Dickerson, and K. S. Matthews. Kinetic studies of inducer binding to lac repressor.operator complex. *J Biol Chem*, 255(21):10115–9, November 1980.

[82] W. H. Orme-Johnson. Molecular basis of biological nitrogen fixation. *Annu Rev Biophys Biophys Chem*, 14:419–59, 1985.

[83] R. N. Thorneley and D. J. Lowe. Nitrogenase of Klebsiella pneumoniae. Kinetics of the dissociation of oxidized iron protein from molybdenum-iron protein: identification of the rate-limiting step for substrate reduction. *Biochem J*, 215(2):393–403, November 1983.

[84] C. Jarzynski. Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters*, 78(14):2690–2693, April 1997.

[85] J. P. Lewis, K. R. Glaesemann, G. A. Voth, J. Fritsch, A. A. Demkov, J. Ortega, and O. F. Sankey. Further developments in the local-orbital density-functional-theory tight-binding method. *Physical Review B*, 64(19):195103, October 2001.

[86] J. P. Lewis, P. Jelínek, J. Ortega, A. A. Demkov, D. G. Trabada, B. Haycock, H. Wang, G. Adams, J. K. Tomfohr, E. Abad, H. Wang, and D. A. Drabold. Advances and applications in the FIREBALL ab initio tight-binding molecular-dynamics formalism. *physica status solidi (b)*, 248(9):1989–2007, September 2011.

[87] J. I. Mendieta-Moreno, R. C. Walker, J. P. Lewis, P. Gomez-Puertas, J. Mendieta, and J. Ortega. fireball/amber: An Efficient Local-Orbital DFT QM/MM Method for Biomolecular Systems. *J Chem Theory Comput*, 10(5):2185–93, May 2014.

[88] J. I. Mendieta-Moreno, I. Marcos-Alcalde, D. G. Trabada, P. Gomez-Puertas, J. Ortega, and J. Mendieta. A Practical Quantum Mechanics Molecular Mechanics Method for the Dynamical Study of Reactions in Biomolecules. *Adv Protein Chem Struct Biol*, 100:67–88, 2015.

[89] P. Hohenberg and W. Kohn. Inhomogeneous Electron Gas. *Physical Review*, 136(3B):B864–B871, November 1964.

[90] W. Kohn and L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A):A1133–A1138, November 1965.

[91] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A*, 38(6):3098–3100, September 1988.

[92] C. Lee, W. Yang, and R. G. Parr. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B*, 37(2):785–789, January 1988.

[93] P. Jelínek, H. Wang, J. P. Lewis, O. F. Sankey, and J. Ortega. Multicenter approach to the exchange-correlation interactions in ab initio tight-binding methods. *Physical Review B*, 71(23):235101, June 2005.

[94] A. A. Demkov, J. Ortega, O. F. Sankey, and M. P. Grumbach. Electronic structure approach for complex silicas. *Physical Review B*, 52(3):1618–1630, 1995.

[95] F. J. García-Vidal, J. Merino, R. Pérez, R. Rincón, J. Ortega, and F. Flores. Density-functional approach to LCAO methods. *Physical Review B*, 50(15):10537–10547, October 1994.

[96] J. I. Mendieta-Moreno. *Simulación de reacciones en biomoléculas con QM/MM.* PhD thesis, Universidad Autónoma de Madrid, 2017.

[97] K. J. Laidler and M. Christine King. Development of transition-state theory. *The Journal of Physical Chemistry*, 87(15):2657–2664, July 1983.

[98] A. D. McNaught and A. Wilkinson. *GoldBook: IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book").* Blackwell Scientific Publications, Oxford, 1997. XML on-line corrected version: http://goldbook.iupac.org (2006-) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins.

[99] D. G. Truhlar. Transition state theory for enzyme kinetics. *Archives of Biochemistry and Biophysics*, 582:10–17, September 2015.

[100] H. Eyring. The Activated Complex in Chemical Reactions. *The Journal of Chemical Physics*, 3(2):107–115, February 1935.

[101] M. G. Evans and M. Polanyi. Some applications of the transition state method to the calculation of reaction velocities, especially in solution. *Transactions of the Faraday Society*, 31(0):875–894, January 1935.

[102] T. Sterling and J. J. Irwin. ZINC 15–Ligand Discovery for Everyone. *J Chem Inf Model*, 55(11):2324–37, November 2015.

[103] D. R. Koes, M. P. Baumgartner, and C. J. Camacho. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model*, 53(8):1893–904, August 2013.

[104] O. Trott and A. J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, January 2010.

[105] J. Baxter. Local Optima Avoidance in Depot Location. *Journal of the Operational Research Society*, 32(9):815–819, September 1981.

[106] A. R. C. Blum, Blesa, M., and Sampels, M. *Hybrid Metaheuristics - An Emerging Approach to Optimization*. Springer-Verlag, 2008.

[107] I. Shcherbakova, S. Mitra, A. Laederach, and M. Brenowitz. Energy barriers, pathways, and dynamics during folding of large, multidomain RNAs. *Curr Opin Chem Biol*, 12(6):655–66, December 2008.

[108] N. Hori, G. Chikenji, R. S. Berry, and S. Takada. Folding energy landscape and network dynamics of small globular proteins. *Proc Natl Acad Sci U S A*, 106(1):73–8, January 2009.

[109] F. Bai, Y. Xu, J. Chen, Q. Liu, J. Gu, X. Wang, J. Ma, H. Li, J. N. Onuchic, and H. Jiang. Free energy landscape for the binding process of Huperzine A to acetylcholinesterase. *Proc Natl Acad Sci U S A*, 110(11):4273–8, March 2013.

[110] F. Martin-Garcia, J. I. Mendieta-Moreno, E. Lopez-Vinas, P. Gomez-Puertas, and J. Mendieta. The Role of Gln61 in HRas GTP hydrolysis: a quantum mechanics/-molecular mechanics study. *Biophys J*, 102(1):152–7, January 2012.

[111] F. Martin-Garcia, J. I. Mendieta-Moreno, I. Marcos-Alcalde, P. Gomez-Puertas, and J. Mendieta. Simulation of catalytic water activation in mitochondrial F1-ATPase using a hybrid quantum mechanics/molecular mechanics approach: an alternative role for beta-Glu 188. *Biochemistry*, 52(5):959–66, February 2013.

[112] J. I. Mendieta-Moreno, D. G. Trabada, J. Mendieta, J. P. Lewis, P. Gomez-Puertas, and J. Ortega. Quantum Mechanics/Molecular Mechanics Free Energy Maps and Nonadiabatic Simulations for a Photochemical Reaction in DNA: Cyclobutane Thymine Dimer. *J Phys Chem Lett*, 7(21):4391–4397, November 2016.

[113] A. Kachmar, M. Carignano, T. Laino, M. Iannuzzi, and J. Hutter. Mapping the Free Energy of Lithium Solvation in the Protic Ionic Liquid Ethylammonuim Nitrate: A Metadynamics Study. *ChemSusChem*, 10(15):3083–3090, August 2017.

[114] A. Kachmar and W. A. Goddard III. The Role of Solvent for Sodium Intercalation into Graphite. *arXiv:1802.06689 [cond-mat, physics:physics]*, February 2018. arXiv: 1802.06689.

[115] Shipman, J. W. *Tkinter Reference: a GUI for Python.* New Mexico Tech Computer Center, Socorro, New Mexico., 2010.

[116] J. D. Hunter. Matplotlib: A 2d Graphics Environment. *Computing in Science Engineering*, 9(3):90–95, May 2007.

[117] S. van der Walt, S. C. Colbert, and G. Varoquaux. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering*, 13(2):22–30, March 2011.

[118] E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numer. Math.*, 1(1):269–271, December 1959.

[119] A. Lengronne, J. McIntyre, Y. Katou, Y. Kanoh, K. P. Hopfner, K. Shirahige, and F. Uhlmann. Establishment of sister chromatid cohesion at the S. cerevisiae replication fork. *Mol Cell*, 23(6):787–99, September 2006.

[120] S. Rankin and D. S. Dawson. Recent advances in cohesin biology. *F1000Res*, 5, 2016.

[121] C. H. Haering and S. Gruber. SnapShot: SMC Protein Complexes Part II. *Cell*, 164(4):818 e1, February 2016.

[122] J. Liu and I. D. Krantz. Cornelia de Lange syndrome, cohesin, and beyond. *Clin Genet*, 76(4):303–14, October 2009.

[123] G. D. Mehta, R. Kumar, S. Srivastava, and S. K. Ghosh. Cohesin: functions beyond sister chromatid cohesion. *FEBS Lett*, 587(15):2299–312, August 2013.

[124] A. L. Sanborn, S. S. Rao, S. C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, 112(47):E6456–65, November 2015.

[125] C. Barrington, R. Finn, and S. Hadjur. Cohesin biology meets the loop extrusion model. *Chromosome Res*, 25(1):51–60, March 2017.

[126] C. A. Brackley, J. Johnson, D. Michieletto, A. N. Morozov, M. Nicodemi, P. R. Cook, and D. Marenduzzo. Extrusion without a motor: a new take on the loop extrusion model of genome organization. *Nucleus*, 9(1):95–103, January 2018.

[127] C. H. Haering and S. Gruber. SnapShot: SMC Protein Complexes Part I. *Cell*, 164(1-2):326–326 e1, January 2016.

[128] F. Uhlmann. SMC complexes: from DNA to chromosomes. *Nat Rev Mol Cell Biol*, 17(7):399–412, July 2016.

[129] S. H. Harvey, M. J. Krien, and M. J. O'Connell. Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. *Genome Biol*, 3(2):REVIEWS3003, 2002.

[130] M. L. Diebold-Durand, H. Lee, L. B. Ruiz Avila, H. Noh, H. C. Shin, H. Im, F. P. Bock, F. Burmann, A. Durand, A. Basfeld, S. Ham, J. Basquin, B. H. Oh,

and S. Gruber. Structure of Full-Length SMC and Rearrangements Required for Chromosome Organization. *Mol Cell*, 67(2):334–347 e5, July 2017.

[131] T. G. Gligoris, J. C. Scheinost, F. Burmann, N. Petela, K. L. Chan, P. Uluocak, F. Beckouet, S. Gruber, K. Nasmyth, and J. Lowe. Closing the cohesin ring: structure and function of its Smc3-kleisin interface. *Science*, 346(6212):963–7, November 2014.

[132] Y. Murayama and F. Uhlmann. DNA Entry into and Exit out of the Cohesin Ring by an Interlocking Gate Mechanism. *Cell*, 163(7):1628–40, December 2015.

[133] C. H. Haering, D. Schoffnegger, T. Nishino, W. Helmhart, K. Nasmyth, and J. Lowe. Structure and stability of cohesin's Smc1-kleisin interaction. *Mol Cell*, 15(6):951–64, September 2004.

[134] P. J. Huis in 't Veld, F. Herzog, R. Ladurner, I. F. Davidson, S. Piric, E. Kreidl, V. Bhaskara, R. Aebersold, and J. M. Peters. Characterization of a DNA exit gate in the human cohesin ring. *Science*, 346(6212):968–72, November 2014.

[135] R. Ladurner, V. Bhaskara, P. J. Huis in 't Veld, I. F. Davidson, E. Kreidl, G. Petzold, and J. M. Peters. Cohesin's ATPase activity couples cohesin loading onto DNA with Smc3 acetylation. *Curr Biol*, 24(19):2228–37, October 2014.

[136] Y. Murayama and F. Uhlmann. Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature*, 505(7483):367–71, January 2014.

[137] P. Arumugam, S. Gruber, K. Tanaka, C. H. Haering, K. Mechtler, and K. Nasmyth. ATP hydrolysis is required for cohesin's association with chromosomes. *Curr Biol*, 13(22):1941–53, November 2003.

[138] S. Weitzer, C. Lehane, and F. Uhlmann. A model for ATP hydrolysis-dependent binding of cohesin to DNA. *Curr Biol*, 13(22):1930–40, November 2003.

[139] A. M. O. Elbatsh, J. H. I. Haarhuis, N. Petela, C. Chapard, A. Fish, P. H. Celie, M. Stadnik, D. Ristic, C. Wyman, R. H. Medema, K. Nasmyth, and B. D. Rowland. Cohesin Releases DNA through Asymmetric ATPase-Driven Ring Opening. *Mol Cell*, 61(4):575–588, February 2016.

[140] F. Beckouet, M. Srinivasan, M. B. Roig, K. L. Chan, J. C. Scheinost, P. Batty, B. Hu, N. Petela, T. Gligoris, A. C. Smith, L. Strmecki, B. D. Rowland, and K. Nasmyth. Releasing Activity Disengages Cohesin's Smc3/Scc1 Interface in a Process Blocked by Acetylation. *Mol Cell*, 61(4):563–574, February 2016.

[141] P. Arumugam, T. Nishino, C. H. Haering, S. Gruber, and K. Nasmyth. Cohesin's ATPase activity is stimulated by the C-terminal Winged-Helix domain of its kleisin subunit. *Curr Biol*, 16(20):1998–2008, October 2006.

[142] C. Gervasini, S. Russo, A. Cereda, I. Parenti, M. Masciadri, J. Azzollini, D. Melis, T. Aravena, B. Doray, A. Ferrarini, L. Garavelli, A. Selicorni, and L. Larizza. Cornelia de Lange individuals with new and recurrent SMC1a mutations enhance delineation of mutation repertoire and phenotypic spectrum. *Am J Med Genet A*, 161A(11):2909–19, November 2013.

[143] L. Mannini, J. Liu, I. D. Krantz, and A. Musio. Spectrum and consequences of

SMC1a mutations: the unexpected involvement of a core component of cohesin in human disease. *Hum Mutat*, 31(1):5–10, January 2010.

[144] J. Liu, R. Feldman, Z. Zhang, M. A. Deardorff, E. V. Haverfield, M. Kaur, J. R. Li, D. Clark, A. D. Kline, D. J. Waggoner, S. Das, L. G. Jackson, and I. D. Krantz. SMC1a expression and mechanism of pathogenicity in probands with X-Linked Cornelia de Lange syndrome. *Hum Mutat*, 30(11):1535–42, November 2009.

[145] L. Mannini, F. Cucco, V. Quarantotti, I. D. Krantz, and A. Musio. Mutation spectrum and genotype-phenotype correlation in Cornelia de Lange syndrome. *Hum Mutat*, 34(12):1589–96, December 2013.

[146] I. Barisic, V. Tokic, M. Loane, F. Bianchi, E. Calzolari, E. Garne, D. Wellesley, and H. Dolk. Descriptive epidemiology of Cornelia de Lange syndrome in Europe. *Am J Med Genet A*, 146A(1):51–9, January 2008.

[147] M. I. Boyle, C. Jespersgaard, K. Brondum-Nielsen, A. M. Bisgaard, and Z. Tumer. Cornelia de Lange syndrome. *Clin Genet*, 88(1):1–12, July 2015.

[148] I. Parenti, M. E. Teresa-Rodrigo, J. Pozojevic, S. Ruiz Gil, I. Bader, D. Braunholz, N. C. Bramswig, C. Gervasini, L. Larizza, L. Pfeiffer, F. Ozkinay, F. Ramos, B. Reiz, O. Rittinger, T. M. Strom, E. Watrin, K. Wendt, D. Wieczorek, B. Wollnik, C. Baquero-Montoya, J. Pie, M. A. Deardorff, G. Gillessen-Kaesbach, and F. J. Kaiser. Mutations in chromatin regulators functionally link Cornelia de Lange syndrome and clinically overlapping phenotypes. *Hum Genet*, 136(3):307–320, March 2017.

[149] M. A. Deardorff, N. J. Porter, and D. W. Christianson. Structural aspects of HDAC8 mechanism and dysfunction in Cornelia de Lange syndrome spectrum disorders. *Protein Sci*, 25(11):1965–1976, November 2016.

[150] K. Nasmyth. Cohesin: a catenase with separate entry and exit gates? *Nat Cell Biol*, 13(10):1170–7, October 2011.

[151] M. Gause, Z. Misulovin, A. Bilyeu, and D. Dorsett. Dosage-sensitive regulation of cohesin chromosome binding and dynamics by Nipped-B, Pds5, and Wapl. *Mol Cell Biol*, 30(20):4940–51, October 2010.

[152] B. Hu, T. Itoh, A. Mishra, Y. Katoh, K. L. Chan, W. Upcher, C. Godlee, M. B. Roig, K. Shirahige, and K. Nasmyth. ATP hydrolysis is required for relocating cohesin from sites occupied by its Scc2/4 loading complex. *Curr Biol*, 21(1):12–24, January 2011.

[153] V. Bajic, B. Spremo-Potparevic, L. Zivkovic, E. R. Isenovic, and T. Arendt. Cohesion and the aneuploid phenotype in Alzheimer's disease: A tale of genome instability. *Neurosci Biobehav Rev*, 55:365–74, August 2015.

[154] S. Cuylen, J. Metz, and C. H. Haering. Condensin structures chromosomal DNA through topological links. *Nat Struct Mol Biol*, 18(8):894–901, July 2011.

[155] C. H. Haering, A. M. Farcas, P. Arumugam, J. Metson, and K. Nasmyth. The cohesin ring concatenates sister DNA molecules. *Nature*, 454(7202):297–301, July 2008.

[156] R. Ciosk, M. Shirayama, A. Shevchenko, T. Tanaka, A. Toth, and K. Nasmyth. Cohesin's binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins. *Mol Cell*, 5(2):243–54, February 2000.

[157] R. Gandhi, P. J. Gillespie, and T. Hirano. Human Wapl is a cohesin-binding protein that promotes sister-chromatid resolution in mitotic prophase. *Curr Biol*, 16(24):2406–17, December 2006.

[158] S. Kueng, B. Hegemann, B. H. Peters, J. J. Lipp, A. Schleiffer, K. Mechtler, and J. M. Peters. Wapl controls the dynamic association of cohesin with chromatin. *Cell*, 127(5):955–67, December 2006.

[159] T. Rolef Ben-Shahar, S. Heeger, C. Lehane, P. East, H. Flynn, M. Skehel, and F. Uhlmann. Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science*, 321(5888):563–6, July 2008.

[160] B. D. Rowland, M. B. Roig, T. Nishino, A. Kurze, P. Uluocak, A. Mishra, F. Beckouet, P. Underwood, J. Metson, R. Imre, K. Mechtler, V. L. Katis, and K. Nasmyth. Building sister chromatid cohesion: smc3 acetylation counteracts an antiestablishment activity. *Mol Cell*, 33(6):763–74, March 2009.

[161] E. Unal, J. M. Heidinger-Pauli, W. Kim, V. Guacci, I. Onn, S. P. Gygi, and D. E. Koshland. A molecular determinant for the establishment of sister chromatid cohesion. *Science*, 321(5888):566–9, July 2008.

[162] J. Zhang, X. Shi, Y. Li, B. J. Kim, J. Jia, Z. Huang, T. Yang, X. Fu, S. Y. Jung, Y. Wang, P. Zhang, S. T. Kim, X. Pan, and J. Qin. Acetylation of Smc3 by Eco1 is required for S phase sister chromatid cohesion in both human and yeast. *Mol Cell*, 31(1):143–51, July 2008.

[163] K. L. Chan, M. B. Roig, B. Hu, F. Beckouet, J. Metson, and K. Nasmyth. Cohesin's DNA exit gate is distinct from its entrance gate and is regulated by acetylation. *Cell*, 150(5):961–74, August 2012.

[164] L. Lopez-Serra, A. Lengronne, V. Borges, G. Kelly, and F. Uhlmann. Budding yeast Wapl controls sister chromatid cohesion maintenance and chromosome condensation. *Curr Biol*, 23(1):64–9, January 2013.

[165] R. G. Huber, I. Kulemzina, K. Ang, A. P. Chavda, S. Suranthran, J. T. Teh, D. Kenanov, G. Liu, G. Rancati, R. Szmyd, P. Kaldis, P. J. Bond, and D. Ivanov. Impairing Cohesin Smc1/3 Head Engagement Compensates for the Lack of Eco1 Function. *Structure*, 24(11):1991–1999, November 2016.

[166] S. Hayashi, H. Ueno, A. R. Shaikh, M. Umemura, M. Kamiya, Y. Ito, M. Ikeguchi, Y. Komoriya, R. Iino, and H. Noji. Molecular mechanism of ATP hydrolysis in F1-ATPase revealed by molecular simulations and single-molecule observations. *J Am Chem Soc*, 134(20):8447–54, May 2012.

[167] H. Wackerhage, U. Hoffmann, D. Essfeld, D. Leyk, K. Mueller, and J. Zange. Recovery of free ADP, Pi, and free energy of ATP hydrolysis in human skeletal muscle. *J Appl Physiol (1985)*, 85(6):2140–5, December 1998.

[168] S. A. Forbes, D. Beare, P. Gunasekaran, K. Leung, N. Bindal, H. Boutselakis, M. Ding, S. Bamford, C. Cole, S. Ward, C. Y. Kok, M. Jia, T. De, J. W. Teague,

M. R. Stratton, U. McDermott, and P. J. Campbell. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, 43(Database issue):D805–11, January 2015.

[169] A. D. Kline, I. D. Krantz, A. Sommer, M. Kliewer, L. G. Jackson, D. R. FitzPatrick, A. V. Levin, and A. Selicorni. Cornelia de Lange syndrome: clinical review, diagnostic and scoring systems, and anticipatory guidance. *Am J Med Genet A*, 143A(12):1287–96, June 2007.

[170] A. D. Kline, M. Grados, P. Sponseller, H. P. Levy, N. Blagowidow, C. Schoedel, J. Rampolla, D. K. Clemens, I. Krantz, A. Kimball, C. Pichard, and D. Tuchman. Natural history of aging in Cornelia de Lange syndrome. *Am J Med Genet C Semin Med Genet*, 145C(3):248–60, August 2007.

[171] M. Ansari, G. Poke, Q. Ferry, K. Williamson, R. Aldridge, A. M. Meynert, H. Bengani, C. Y. Chan, H. Kayserili, S. Avci, R. C. Hennekam, A. K. Lampe, E. Redeker, T. Homfray, A. Ross, M. Falkenberg Smeland, S. Mansour, M. J. Parker, J. A. Cook, M. Splitt, R. B. Fisher, A. Fryer, A. C. Magee, A. Wilkie, A. Barnicoat, A. F. Brady, N. S. Cooper, C. Mercer, C. Deshpande, C. P. Bennett, D. T. Pilz, D. Ruddy, D. Cilliers, D. S. Johnson, D. Josifova, E. Rosser, E. M. Thompson, E. Wakeling, E. Kinning, F. Stewart, F. Flinter, K. M. Girisha, H. Cox, H. V. Firth, H. Kingston, J. S. Wee, J. A. Hurst, J. Clayton-Smith, J. Tolmie, J. Vogt, K. Tatton-Brown, K. Chandler, K. Prescott, L. Wilson, M. Behnam, M. McEntagart, R. Davidson, S. A. Lynch, S. Sisodiya, S. G. Mehta, S. A. McKee, S. Mohammed, S. Holden, S. M. Park, S. E. Holder, V. Harrison, V. McConnell, W. K. Lam, A. J. Green, D. Donnai, M. Bitner-Glindzicz, D. E. Donnelly, C. Nellaker, M. S. Taylor, and D. R. FitzPatrick. Genetic heterogeneity in Cornelia de Lange syndrome (CdLS) and CdLS-like phenotypes with observed and predicted levels of mosaicism. *J Med Genet*, 51(10):659–68, October 2014.

[172] D. Mouradov, C. Sloggett, R. N. Jorissen, C. G. Love, S. Li, A. W. Burgess, D. Arango, R. L. Strausberg, D. Buchanan, S. Wormald, L. O'Connor, J. L. Wilding, D. Bicknell, I. P. Tomlinson, W. F. Bodmer, J. M. Mariadason, and O. M. Sieber. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res*, 74(12):3238–47, June 2014.

[173] I. Kulemzina, K. Ang, X. Zhao, J. T. Teh, V. Verma, S. Suranthran, A. P. Chavda, R. G. Huber, B. Eisenhaber, F. Eisenhaber, J. Yan, and D. Ivanov. A Reversible Association between Smc Coiled Coils Is Regulated by Lysine Acetylation and Is Required for Cohesin Association with the DNA. *Mol Cell*, 63(6):1044–54, September 2016.

[174] W. C. Chao, B. O. Wade, C. Bouchoux, A. W. Jones, A. G. Purkiss, S. Federico, N. O'Reilly, A. P. Snijders, F. Uhlmann, and M. R. Singleton. Structural Basis of Eco1-Mediated Cohesin Acetylation. *Sci Rep*, 7:44313, March 2017.

[175] Y. Liu, S. Sung, Y. Kim, F. Li, G. Gwon, A. Jo, A. K. Kim, T. Kim, O. K. Song, S. E. Lee, and Y. Cho. ATP-dependent DNA binding, unwinding, and resection by the Mre11/Rad50 complex. *EMBO J*, 35(7):743–58, April 2016.

[176] H. Schuler and C. Sjogren. DNA binding to SMC ATPases-trapped for release. *EMBO J*, 35(7):703–5, April 2016.

[177] F. U. Seifert, K. Lammens, G. Stoehr, B. Kessler, and K. P. Hopfner. Structural mechanism of ATP-dependent DNA binding and DNA end bridging by eukaryotic Rad50. *EMBO J*, 35(7):759–72, April 2016.

[178] A. Lammens, A. Schele, and K. P. Hopfner. Structural biochemistry of ATP-driven dimerization and DNA-stimulated activation of SMC ATPases. *Curr Biol*, 14(19):1778–82, October 2004.

[179] K. P. Hopfner. Invited review: Architectures and mechanisms of ATP binding cassette proteins. *Biopolymers*, 105(8):492–504, August 2016.

[180] K. Kamada, M. Su'etsugu, H. Takada, M. Miyata, and T. Hirano. Overall Shapes of the SMC-ScpAB Complex Are Determined by Balance between Constraint and Relaxation of Its Structural Parts. *Structure*, 25(4):603–616 e4, April 2017.

[181] M. De Vivo. Bridging quantum mechanics and structure-based drug design. *Front Biosci (Landmark Ed)*, 16:1619–33, January 2011.

[182] V. L. Schramm. Transition States, analogues, and drug development. *ACS Chem Biol*, 8(1):71–81, January 2013.

[183] V. L. Schramm. Transition States and transition state analogue interactions with enzymes. *Acc Chem Res*, 48(4):1032–9, April 2015.

[184] M. T. Hons, P. J. Huis In 't Veld, J. Kaesler, P. Rombaut, A. Schleiffer, F. Herzog, H. Stark, and J. M. Peters. Topology and structure of an engineered human cohesin complex bound to Pds5b. *Nat Commun*, 7:12523, August 2016.

[185] J.P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.

[186] S. Miyamoto and P. A. Kollman. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry*, 13(8):952–962, October 1992.

[187] S. Kalyaanamoorthy and Y. P. Chen. A steered molecular dynamics mediated hit discovery for histone deacetylases. *Phys Chem Chem Phys*, 16(8):3777–91, February 2014.

[188] L. S. Cheung, D. J. Shea, N. Nicholes, A. Date, M. Ostermeier, and K. Konstantopoulos. Characterization of monobody scaffold interactions with ligand via force spectroscopy and steered molecular dynamics. *Sci Rep*, 5:8247, February 2015.

[189] J. Sadowski, J. Gasteiger, and G. Klebe. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-ray Structures. *Journal of Chemical Information and Computer Sciences*, 34(4):1000–1008, July 1994.

[190] C.H. Schwab. Conformations and 3d pharmacophore searching. *Drug Discovery Today: Technologies*, 7(4):e245–e253, December 2010.

[191] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–91, December 2009.

[192] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison. Open Babel: An open chemical toolbox. *J Cheminform*, 3:33, October 2011.

[193] M. Conese and A. Liso. Cohesin complex is a major player on the stage of leukemogenesis. *Stem Cell Investig*, 3:18, 2016.

Appendices

A Original paper: "MEPSA: minimum energy pathway analysis for energy landscapes"

Bioinformatics, 31(23), 2015, 3853–3855
doi: 10.1093/bioinformatics/btv453
Advance Access Publication Date: 31 July 2015
Applications Note

OXFORD

Structural bioinformatics

MEPSA: minimum energy pathway analysis for energy landscapes

Iñigo Marcos-Alcalde[1], Javier Setoain[2], Jesús I. Mendieta-Moreno[1,3], Jesús Mendieta[1,4] and Paulino Gómez-Puertas[1,*]

[1]Molecular Modelling Group, CBMSO (CSIC-UAM), ES-28049 Madrid, Spain, [2]Departamento de Arquitectura de Computadores y Automática, UCM, ES-28040 Madrid, Spain, [3]Departamento de Física Teórica de la Materia Condensada and Condensed Matter Physics Center (IFIMAC), UAM and [4]Biomol-Informatics SL, Campus UAM, ES-28049 Madrid, Spain

*To whom correspondence should be addressed.
Associate Editor: Anna Tramontano

Received on June 12, 2015; revised on July 20, 2015; accepted on July 24, 2015

Abstract

Summary: From conformational studies to atomistic descriptions of enzymatic reactions, potential and free energy landscapes can be used to describe biomolecular systems in detail. However, extracting the relevant data of complex 3D energy surfaces can sometimes be laborious. In this article, we present MEPSA (Minimum Energy Path Surface Analysis), a cross-platform user friendly tool for the analysis of energy landscapes from a transition state theory perspective. Some of its most relevant features are: identification of all the barriers and minima of the landscape at once, description of maxima edge profiles, detection of the lowest energy path connecting two minima and generation of transition state theory diagrams along these paths. In addition to a built-in plotting system, MEPSA can save most of the generated data into easily parseable text files, allowing more versatile uses of MEPSA's output such as the generation of molecular dynamics restraints from a calculated path.
Availability and implementation: MEPSA is freely available (under GPLv3 license) at: http://bioweb.cbm.uam.es/software/MEPSA/
Contact: pagomez@cbm.csic.es
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The development of efficient conformational space sampling methodologies, coupled with a dramatic increase in the computational capacities and capabilities, has made the calculation of energy landscapes significantly more accessible (Bernardi *et al.*, 2015; Mendieta-Moreno *et al.*, 2014).

A 3D energy surface may provide a lot of detailed information about biomolecular processes such as protein or nucleotide folding (Hori *et al.*, 2009; Shcherbakova *et al.*, 2008), ligand binding (Bai *et al.*, 2013) or enzymatic reactions (Mendieta-Moreno *et al.*, 2015); but the extraction and analysis of such data are often cumbersome tasks. MEPSA (Minimum Energy Path Surface Analysis) provides a GUI-based tool to analyse these landscapes from a transition state theory point of view, making the analysis of 3D energy landscapes agile.

2 MEPSA software

MEPSA is an open-source program written in Python (compatible with both Python 2.7.x and 3.4.x) that describes the connectivity of the minima (called nodes in MEPSA) present in a given energy landscape (called map in MEPSA). The graphic interface is built with TKinter (Shipman, 2010), plots are drawn using Matplotlib (Hunter, 2007) and NumPy (Van der Walt *et al.*, 2011) is employed for the calculations.

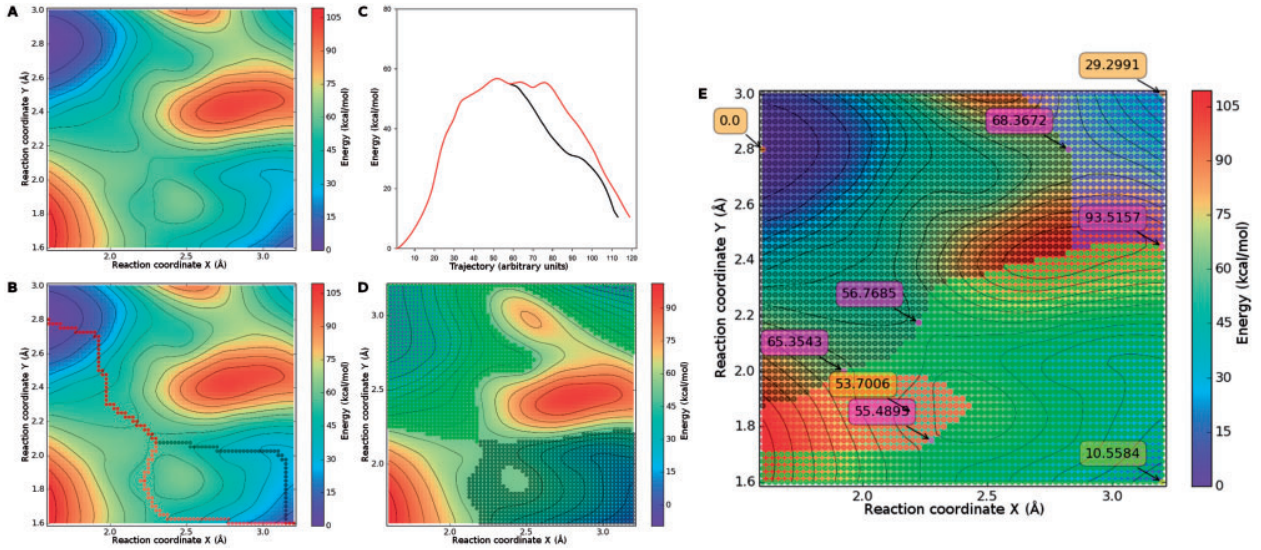


Fig. 1. Some results of an analysis with MEPSA. (**A**) Plot of the test map. (**B**) Comparison of two alternative paths connecting the origin [1.6, 2.8] and the target [3.2, 1.6] nodes. (**C**) Energy profile of both paths. (**D**) Well sampling analysis showing the points sampled to reach the barrier from the origin (green) and the target (black). (**E**) Global connectivity analysis output, highlighting the energy values of the points of interest (minima in orange and saddle points in purple)

MEPSA supports three column formatted plain text files as input and can save most of the generated data in column formatted plain text files as well. There is no restriction related to technique with which the map has been generated, apart from being rectangular and uniformly distributed (e.g. calculation of the minimum elevation path connecting Geneva and Turin is included in Supplementary Material). In addition, any plot obtained (e.g. Fig. 1) can be directly saved in many different image formats (png, eps, pdf . . .).

The GUI structure consists of one main window to load, unload and plot maps and two secondary windows to manage specific tasks. The 'connectivity analyses' window gives access to the 'global connectivity analysis', 'path generator' and 'well sampling analysis' tools; and the 'map editor' window gives access to the 'modify map', 'smooth map' and 'invert map' functionalities.

'Global connectivity analysis' simultaneously samples the whole map, starting from every node and iteratively occupying the lowest energy points available. The node where each particular propagation comes from is kept in memory in order to assign a domain to each node. As the lowest energy points in the borders between domains are necessarily the barriers connecting the nodes, all the minima and barriers of the map are identified at once (Fig. 1E).

'Path generator' detects the lowest energy path connecting two points (named origin and target). The resulting trajectory only depends on the points selected and not on the path direction. MEPSA offers two sampling modes: 'global' and 'node by node'. 'Global' mode uses an approach similar to Dijkstra's algorithm (Dijkstra, 1959), with small differences in the sampling and the trace back. The algorithm samples the map from the origin point, propagating to the lowest energy points available on each iteration, until the target point is reached. The iteration in which each node has been occupied is stored in memory and, after the target node is reached, a trace-back is performed from the target, iteratively selecting the points with lowest iteration counters. 'Node by node' mode uses a 'global' mode sampling to define the order in which the nodes are visited, then performing a series of runs connecting all the consecutive nodes in a pairwise fashion. This way the path is forced to pass through every node whose domain is crossed by the shortest lowest energy path. The paths generated with the 'path generator' can be stored in memory using the 'path stack', enabling the comparison of several paths at once (Fig. 1B and C), even if those were generated using different maps.

'Well sampling analysis' determines the area of the map that has to be sampled from the origin node to reach the closest barrier to the target node and *vice versa* (Fig. 1D).

The data obtained can be plotted in several ways, the path trajectories can be smoothed and most of the data generated (even the 'path stack' as a whole) can be stored into text files to be plotted or analysed with other software (e.g. parsing path files in order to generate molecular dynamics restraints, see Supplementary Material).

In the 'map editor window', 'modify map' performs simple modifications of the energy values in a defined region of the map, which can be useful, e.g. to block favourable paths in order to evaluate alternative ones (Fig. 1B and C); 'smooth map' applies a simple running average smoothing to the map, allowing the user to remove unwanted local minima in noisy maps; and 'invert map' changes the sign of the energy values, enabling the characterization of maxima edge profiles (see Supplementary Material).

Acknowledgements

Support from the 'Fundación Severo Ochoa' and the 'Centro de Computación Científica CCC-UAM' is gratefully acknowledged.

Funding

Grant IPT2011-0964-900000 (Government of Spain). Work at Biomol-Informatics was financed by the European Social Fund.

Conflict of Interest: none declared.

References

Bai,F. *et al.* (2013) Free energy landscape for the binding process of Huperzine A to acetylcholinesterase. *Proc. Natl. Acad. Sci. USA.*, **110**, 4273–4278.

Bernardi,R.C. *et al.* (2015) Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta*, **1850**, 872–877.

Dijkstra,E.W. (1959) A note on two problems in connection with graphs. *Numerische Mathematik*, **1**, 269–271.

Hori,N. *et al*. (2009) Folding energy landscape and network dynamics of small globular proteins. *Proc. Natl. Acad. Sci. USA*., **106**, 73–78.

Hunter,J.D. (2007) Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.*, **9**, 22–30.

Mendieta-Moreno,J.I. *et al*. (2014) FIREBALL/AMBER: an efficient local-orbital DFT QM/MM method for biomolecular systems. *J. Chem. Theor. Comput.*, **10**, 2185–2193.

Mendieta-Moreno,J.I. *et al*. (2015) A practical quantum mechanics molecular mechanics method for the dynamical study of reactions in biomolecules. *Adv. Protein Chem. Struct. Biol.*, **100**, 67–88.

Shcherbakova,I. *et al*. (2008) Energy barriers, pathways and dynamics during folding of large, multi-domain RNAs. *Curr. Opin. Chem. Biol.* **12**, 655–666.

Shipman,J.W. (2010) *Tkinter Reference: a GUI for Python*. New Mexico Tech Computer Center, Socorro, New Mexico.

Van der Walt,S. *et al*. (2011) The NumPy Array: a structure for efficient numerical computation. *Comput. Sci. Eng.*, **13**, 22–30.

B | Original paper: "Two-step ATP-driven opening of cohesin head"

SCIENTIFIC REPORTS

OPEN

Two-step ATP-driven opening of cohesin head

Íñigo Marcos-Alcalde[1], Jesús I. Mendieta-Moreno[1,2], Beatriz Puisac[3], María Concepción Gil-Rodríguez[3], María Hernández-Marcos[3], Diego Soler-Polo[2], Feliciano J. Ramos[3], José Ortega[2], Juan Pié[3], Jesús Mendieta[1,2,4] & Paulino Gómez-Puertas[1]

Received: 5 January 2017
Accepted: 24 April 2017
Published online: 12 June 2017

The cohesin ring is a protein complex composed of four core subunits: Smc1A, Smc3, Rad21 and Stag1/2. It is involved in chromosome segregation, DNA repair, chromatin organization and transcription regulation. Opening of the ring occurs at the "head" structure, formed of the ATPase domains of Smc1A and Smc3 and Rad21. We investigate the mechanisms of the cohesin ring opening using techniques of free molecular dynamics (MD), steered MD and quantum mechanics/molecular mechanics MD (QM/MM MD). The study allows the thorough analysis of the opening events at the atomic scale: i) ATP hydrolysis at the Smc1A site, evaluating the role of the carboxy-terminal domain of Rad21 in the process; ii) the activation of the Smc3 site potentially mediated by the movement of specific amino acids; and iii) opening of the head domains after the two ATP hydrolysis events. Our study suggests that the cohesin ring opening is triggered by a sequential activation of the ATP sites in which ATP hydrolysis at the Smc1A site induces ATPase activity at the Smc3 site. Our analysis also provides an explanation for the effect of pathogenic variants related to cohesinopathies and cancer.

Maintenance of the integrity of genomic information is a supreme requirement for all living organisms. In cells, the DNA molecule containing such information is structurally organized in chromosomes, arranged with a number of different protein macromolecular complexes. They are devoted to a variety of functions, from the scaffolding of the chromosomal building to the regulation of the gene expression. The cohesin ring is one of these complexes, an essential nano-machine, powered by ATP hydrolysis, that is capable of encircling DNA strands.

The human cohesin ring is a highly conserved multi-protein structure composed of four major subunits: Smc1A, Smc3, Rad21, and Stag1/2, although only the first three are essential to form molecular rings[1–4]. A heterodimer of Smc1A and Smc3 subunits forms a coiled-coil-structured ring with an ATPase "head" and a "hinge" domain. The C-terminal domain of Rad21 binds to the Smc1A head[5] while its N-terminal domain binds to the proximal coiled-coil segment of Smc3[6,7]. The Smc subunits belong to a family of proteins involved in a large variety of functions related to chromosome structure: chromosome segregation during mitosis and meiosis, DNA repair through homologous recombination, organization of the chromatin during interphase and transcription regulation[3,8–10].

Defects in the cohesin ring have been related to genetic disorders, known as cohesinopathies, such as Cornelia de Lange Syndrome (CdLS), Roberts Syndrome, Warsaw Breakage Syndrome, CAID Syndrome and CHOPS Syndrome[11–16], as well as to several types of cancer[17–25]. Defects in proteins that regulate cohesin function have also been related to aneuploidy in neurons, which is a relevant factor in the development of Alzheimer disease[26].

ATP hydrolysis is required for both DNA loading[27–30] and unloading[31–33]. An Smc head heterodimer has to be formed, sandwiching the ATP molecules, prior to the hydrolysis event[4,5,34]. Although yeast cohesin Smc heads can interact in the absence of Scc1 (the yeast orthologue of human Rad21)[30], interaction with the Scc1 subunit, in particular with its C-terminal domain, stimulates ATP hydrolysis[27,35,36]. The DNA binding-mediated ATPase activity in Smc heads is regulated by the acetylation of several Lys residues located both in the Smc head and in the coiled coils[2,32,33,37,38].

[1]Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), 28049, Madrid, Spain. [2]Departamento de Física Teórica de la Materia Condensada and Condensed Matter Physics Center (IFIMAC), Universidad Autónoma de Madrid, 28049, Madrid, Spain. [3]Unidad de Genética Clínica y Genómica Funcional, Departamento de Farmacología-Fisiología y Departamento de Pediatría, Hospital Clínico Universitario "Lozano Blesa", Facultad de Medicina, Universidad de Zaragoza, ISS-Aragon and CIBERER-GCV02, 50009, Zaragoza, Spain. [4]Departamento de Biotecnología, Universidad Francisco de Vitoria, Pozuelo de Alarcón, 28223, Madrid, Spain. Jesús Mendieta and Paulino Gómez-Puertas contributed equally to this work. Correspondence and requests for materials should be addressed to P.G. (email: pagomez@cbm.csic.es)

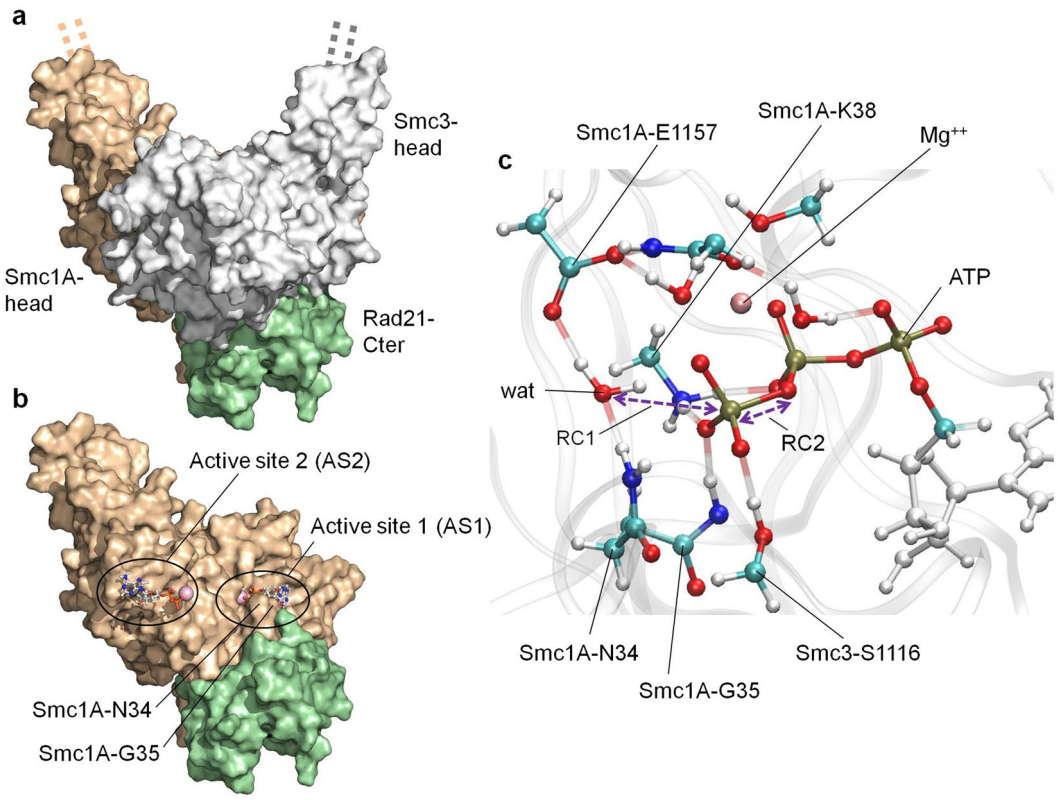


Figure 1. Model overview and description of the QM region. (**a**) Overview of the structural model of the complex formed by the human Smc1A-head (brown), Smc3-head (grey) and Rad21-Cter (green) domains. The dashed lines indicate the direction along which the un-modelled coiled coils would extend towards the hinge domain. (**b**) Location of active site 1 (AS1) and active site 2 (AS2). The Smc1A-head (brown) and Rad21-Cter (green) domains are shown, while the Smc3-head domain is not represented to reveal the location of the active sites. The location of the Smc1A-N34 and Smc1A-G35 residues is indicated. (**c**) QM region of AS1. The atoms in the QM region of the QM/MM MD simulations of AS1 are represented by coloured ball and sticks. The MM regions of the ATP (white ball and sticks) and protein backbone (transparent grey ribbons) are shown. The positions of the catalytic water (wat), residues Smc1A-N34, Smc1A-G35, Smc1A-K38, Smc1A-E1157 and Smc3-S1116, magnesium ion (Mg^{++}) and ATP molecule are indicated. Reaction coordinates 1 (RC1) and 2 (RC2) are indicated by purple arrows.

As indicated in a recent review[4], several questions related to the structure and function of cohesin ring remain open. These deal with: the precise series of events that lead to loading, entrapment, release and stable cohesion; the exact role of the nucleotide binding domains; and how ATP binding and hydrolysis affect the loading and release processes. In addition to the biochemical studies and the highly valuable information offered by the crystallized structures of the Smc head domains[5,7], it is essential to investigate the dynamic properties at the atomic scale in order to study key aspects of cohesin behaviour as well as to analyse and predict the effect of mutations.

In addition, recent studies[18,39,40] indicate that over-expression of the Smc1A protein can play an important oncogenic role in prostate cancer and colorectal cancer. This suggests that this protein is a promising target for anti-tumour drugs. Detailed study, at the structural and quantitative level, of the transition states as rate-limiting steps in the processes of ATP hydrolysis and the opening of Smc heads is of crucial importance for future rational drug design[41–43].

To assess the dynamics of ATP hydrolysis in the cohesin ATPase head heterodimer and its possible effects on the stability of the dimer, we have generated an atomistic model of the ATPase head domains of the human cohesin proteins Smc1A and Smc3 (Smc1A-head and Smc3-head, respectively) bound to the C-terminal domain of human Rad21 (Rad21-Cter) (Fig. 1a). The active site closest to the interface between Smc1A-head and Rad21-Cter, formed by Walker A and Walker B motifs of Smc1A, was labelled as active site 1 (AS1); while the more distant site, formed by Walker A and Walker B motifs of Smc3, was labelled as active site 2 (AS2) (Fig. 1b).

This model mimics the behaviour of the cohesin head, thereby allowing us to investigate and generate functional hypotheses based on detailed analysis of the movement of all the atoms in the head domain; information that cannot be assessed by biochemical assays. In addition, the system enables us to evaluate the role of mutations associated with CdLS and cancer in the functionality of the protein complex.

Results

Rad21 binding induces a rearrangement at active site 1 that allows ATP hydrolysis. It has been reported that the dimerized head domains of Smc1A and Smc3 hydrolyse ATP in the absence of Rad21 with a nearly

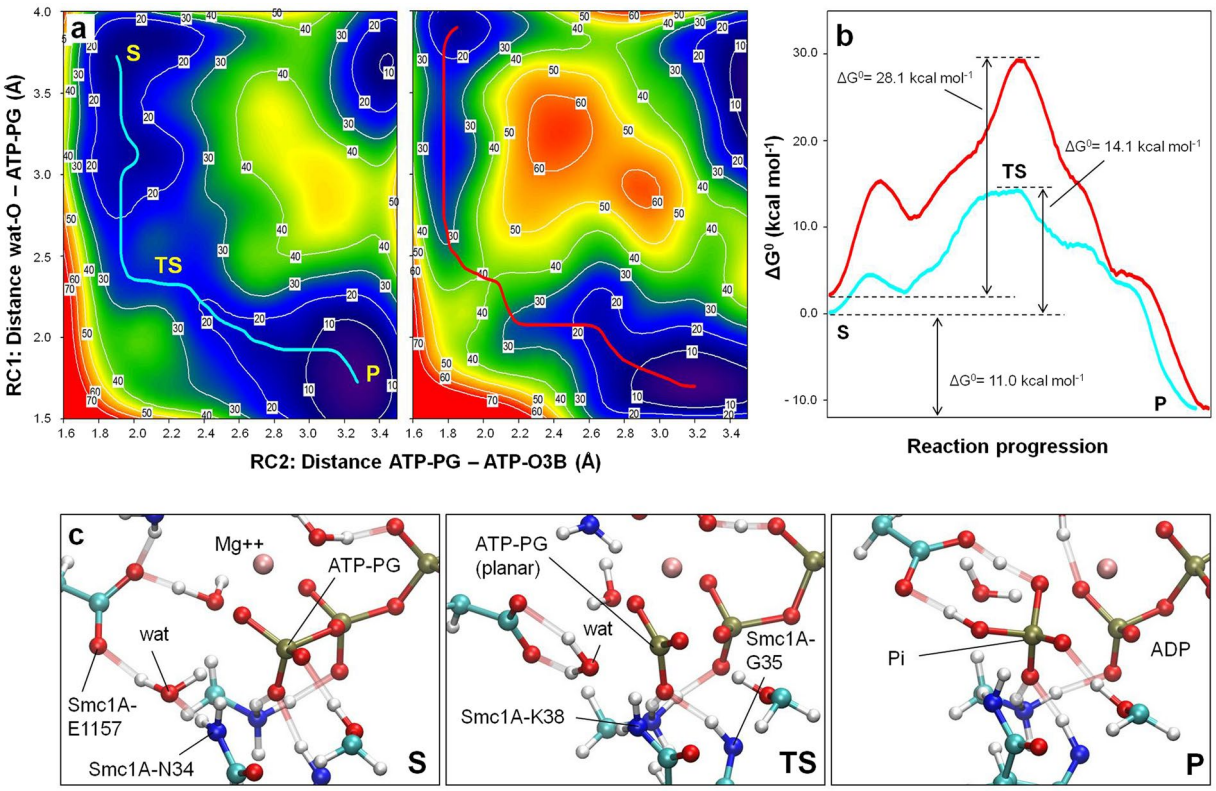


Figure 2. Rad21-Cter allows ATP hydrolysis at AS1. (**a**) Free-energy surfaces (in kcal mol^{-1}) for ATP hydrolysis at AS1 in the presence (left) and absence (right) of Rad21-Cter generated via QM/MM MD simulations. The plot axes represent the reaction coordinates. RC1 (bond to be formed): the distance (in Å) between the oxygen atom of the catalytic water and the phosphorous atom of the ATP molecule γ-phosphate group (distance wat-O - ATP-PG). RC2 (bond to be broken): the distance (in Å) between the phosphorous atom of the ATP molecule γ-phosphate group and the oxygen atom 3 of the ATP β-phosphate group (distance ATP-PG - ATP-O3B). Free-energy data are represented via a colour scale, from lower (blue) to higher (red) values. MEPSA minimum energy paths are shown in cyan (presence of Rad21-Cter) and red (absence of Rad21-Cter). (**b**) Free-energy profiles of the MEPSA minimum-energy paths. The substrate (S), transition state (TS) and product (P) locations are indicated. (**c**) The reference structures of S, TS and P states in the presence of Rad21-Cter are shown. The positions of the catalytic water (wat), residues Smc1A-N34 and Smc1A-E1157, magnesium ion (Mg^{++}), ATP γ-phosphate (ATP-PG), ADP and leaving inorganic phosphate (Pi) are indicated.

undetectable efficiency, but that when they are bound to Rad21-Cter the hydrolysis activity is enhanced[27,35,36]. Using our computational approach, we investigate the hydrolysis reaction at the atomic scale, including the identification of the rate-limiting steps. This analysis provides the explanation for this effect as well as a general model of the movements of all the residues in the complex before and during the ATPase reaction.

In order to evaluate whether the binding of Rad21-Cter directly induces a rearrangement at AS1 that favours ATP hydrolysis, we performed simulations using molecular dynamics and, for the study of the chemical events, our recently developed method for quantum mechanics/molecular mechanics - molecular dynamics (QM/MM MD): Fireball/Amber[44,45]. This fast and accurate method, combining techniques developed in the areas of computational biology (Amber[46]) and condensed matter physics (Fireball[47,48]), permits the generation of 2D free-energy maps of enzymatic reactions without *a priori* determination of the reaction paths. In the present case, we simulated ATP hydrolysis at AS1 through the generation of two equivalent systems of the Smc1A-head/Smc3-head dimer, in the presence and absence of Rad21-Cter. Both systems were stabilized with 40 ns of free molecular dynamics (MD) simulations performed with the Amber14 MD package[46] prior to 150 ps of QM/MM MD stabilization performed using Fireball/Amber. The QM region (Fig. 1c) was formed of the tri-phosphate moiety of the ATP, the magnesium ion, water molecules and side chains present in the coordination sphere of magnesium, the catalytic water molecule and the side chains of Smc1A-N34, Smc1A-G35, Smc1A-K38, Smc1A-E1157 and Smc3-S1116. The MM region comprised the rest of the atoms present in the protein complex and the solvent.

The two free-energy (ΔG°) surfaces were then sampled with Fireball/Amber along two reaction coordinates: reaction coordinate 1 (RC1, the bond to be formed) was the distance between the oxygen atom of the catalytic water and the phosphorous atom of the γ-phosphate group of the ATP molecule while reaction coordinate 2 (RC2, the bond to be broken) was the distance between the γ-phosphate group of the ATP molecule and oxygen atom 3 of the beta phosphate group of ATP (purple arrows in Fig. 1c). The free-energy maps resulting from this sampling (7.6×10^6 conformations and their corresponding total-energies for each map) are depicted in Fig. 2a.

In each case, the minimum energy pathway along the calculated free-energy surfaces (cyan and red lines on the maps in Fig. 2a) was calculated using the MEPSA algorithm[49], by extracting the minimum free-energy path of the reaction from the substrate S (the initial ATP molecule) to the product P (final ADP molecule plus inorganic phosphate group), as represented in Fig. 2b. A detailed explanation of the key reaction steps as well as a video sequence of the whole reaction at AS1 can be found in the Supplementary Information (Supplementary Fig. 1 and Supplementary Video 1).

Briefly, the progression of the ATPase reaction at AS1 in the presence of Rad21-Cter (cyan line in Fig. 2a and b) indicates that the residue Smc1A-N34 plays a crucial role in the entrance of the catalytic water molecule into the active site and its stabilization (Fig. 2c, left), via its coordinated role with the catalytic residue Smc1A-E1157. Smc1A-N34 is maintained in its position by the interaction between Rad21-K605 and Smc1A-G35. The planar structure of the γ-phosphate in the transition state of the reaction (Fig. 2c, centre) is mainly stabilized by Smc1A-K38 and Smc1A-G35 (the latter is also maintained in position by its interaction with Rad21-K605). When the same simulation was run in the absence of Rad21-Cter (red line in Fig. 2a and b), clear differences were observed in the free-energy pathway, which showed higher values during the process of catalytic water accommodation as well as in the transition state (first and second peaks in Fig. 2b). A description of some reaction features at AS1 in presence and absence of Rad21-Cter can be found in the Supplementary Information (Supplementary Figs 2, 3 and 4). The total difference in the free-energy barrier between the two analysed situations was 14.0 kcal mol^{-1} ($\Delta G°$ values of 28.1 kcal mol^{-1} and 14.1 kcal mol^{-1} in the absence or presence of Rad21-Cter respectively). In contrast, the $\Delta G°$ values of the structures at the beginning (S) and the end (P) of the reaction were almost equivalent in the two situations. This indicates that, regarding ATP hydrolysis at AS1, the main effect of being bound to Rad21-Cter is the reduction of the free-energy barrier. This is in agreement with the experimentally observed fact that the presence of Rad21-Cter allows ATP hydrolysis[27,35,36], lowering the barrier to a calculated $\Delta G°$ value close to the range of the experimental free-energy barrier measured for other ATPases, as the F_1-ATPase (12.9–13.4 kcal mol^{-1})[50].

All these results indicate that the binding of Rad21-Cter to the Smc1A-head/Smc3-head dimer induces a rearrangement at AS1 that both facilitates the entrance of the catalytic water molecule into the active site and reduces the energy barrier associated with the transition state.

ATP hydrolysis at active site 1 induces the activation of site 2. Once the ATPase reaction at AS1 was complete, and the site was occupied by the resulting ADP molecule, our next step was to study the effect of this substitution on the structure of the Smc1A-head/Smc3-head/Rad21-Cter complex.

To perform the analysis, the system was subjected to 150 ns of free MD simulation in the presence of an ADP molecule at AS1, while at the same time maintaining an ATP molecule at AS2 (condition AS1-ADP/AS2-ATP). As a control, the same system, but containing ATP at both sites (condition AS1-ATP/AS2-ATP), was subjected to an equivalent simulation. Throughout both trajectories, the movements of residues around the two active centres were monitored for any conformational change that could affect the activity of AS2. Notably, after 120 ns of free MD simulation in the AS1-ADP/AS2-ATP condition, the side chain of an apparently non-related residue, Smc1A-K1120, moved close to the AS2 catalytic water molecule and remained in its new location in a stable conformation (Fig. 3a and Supplementary Video 2). In this condition, the distance between the apical nitrogen atom of the side chain of Smc1A-K1120 and the oxygen atom of the AS2 catalytic water was stabilized at a value of 2.5 Å, in contrast to the value of around 7.9 Å in the AS1-ATP/AS2-ATP condition (Fig. 3b). In the AS1-ADP/AS2-ATP condition, the interaction between Smc1A-K1120 and the catalytic water molecule was found to be stabilized by the formation of a hydrogen bond.

The presence of Smc1A-K1120 in this new position is predicted to dramatically affect the distribution of charges at AS2. To evaluate the extent of this effect, we analysed the ATPase reaction at AS2 in both conditions (AS1-ADP/AS2-ATP and the control AS1-ATP/AS2-ATP) in detail using Fireball/Amber. The initial structures used for both QM/MM MD simulations were the final structures of the 150 ns long free MD simulations shown in Fig. 3a. The QM region (Fig. 3b) was formed of the tri-phosphate moiety of the ATP, the magnesium ion, water molecules and side chains present in the coordination sphere of magnesium, the catalytic water molecule and the side chains of Smc3-K38, Smc3-E1144 and Smc1A-K1120; with the MM region comprising the other atoms in the complex and solvent.

As the entrance of a lysine side chain in close proximity to the catalytic water at AS2 was expected to have a substantial effect on the ATPase activity, the free-energy profiles were initially sampled along a single reaction coordinate, instead of two. The reaction coordinate selected (RC, purple arrow in Fig. 3b) was that corresponding to the bond to be formed: the distance between the oxygen atom of the catalytic water and the phosphorous atom of the γ-phosphate group of the ATP molecule. As expected, 1D free-energy sampling was sensitive enough to detect a remarkable difference between the two conditions. The resulting free-energy profiles for each reaction are depicted in Fig. 3c. A detailed explanation of the key reaction steps as well as a video sequence of the whole reaction at AS2 can be found in the Supplementary Information (Supplementary Figs 3 and 5 and Supplementary Video 3). This QM/MM MD analysis of the reaction revealed the presence of an intermediate transition state during the positioning of the catalytic water molecule, as well as a main transition state corresponding to the planar configuration of the γ-phosphate of the ATP.

Comparison of the 1D free-energy profiles of ATP hydrolysis at AS2 under both conditions showed a reduction of 38.0 kcal mol^{-1} in the energy barrier in the AS1-ADP/AS2-ATP condition compared to the control, with a final value for the reaction barrier of 14.2 kcal mol^{-1} (Fig. 3c). This result reveals that the change in location of Smc1A-K1120, as a result of the presence of ADP at AS1, strongly stimulates the ATPase activity at AS2. In short, our results show that ATP hydrolysis at AS1 induces ATPase activity at AS2 and explain the atomic mechanism for this effect.

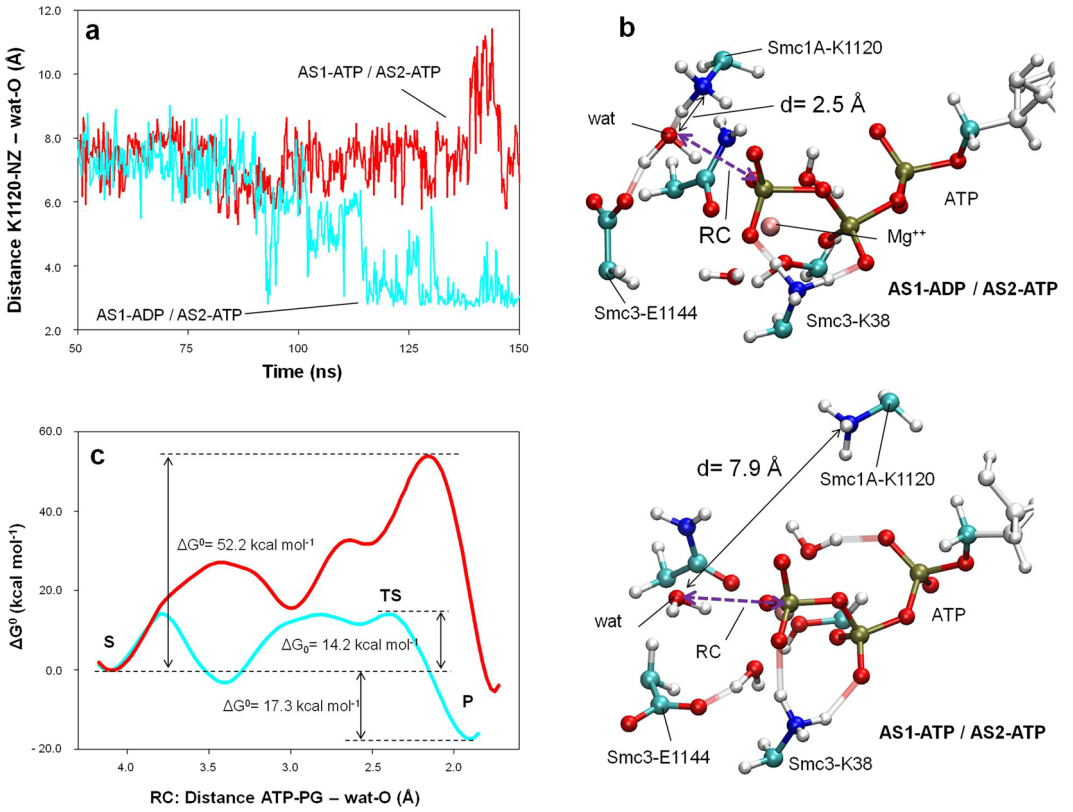


Figure 3. ATP hydrolysis at AS1 activates hydrolysis at AS2. (**a**) Evolution of the distance between the oxygen atom of the catalytic water in AS2 and the ε-amino group of the Smc1A-K1120 residue (distance K1120-NZ - wat-O) prior (red) and after (cyan) ATP hydrolysis at AS1. (**b**) AS2 activation and QM region description. The atoms in the QM region of the QM/MM MD simulations of AS2 are represented by coloured ball and sticks. Part of the MM region of the ATP is shown (white ball and sticks). The positions of the catalytic water (wat), Smc3-K38, Smc3-E1144 and Smc1A-K1120 residues, and the ATP molecule are indicated for both inactive (AS1-ATP/AS2-ATP, red line) and active (AS1-ADP/AS2-ATP, cyan line) AS2 configurations. The distance between the catalytic water and ε-amino group of the Smc1A-K1120 residue is indicated by a black arrow. The reaction coordinate (RC) is indicated by a purple arrow. (**c**) Free-energy (kcal mol^{-1}) profiles generated via QM/MM MD simulations of AS2 in both inactive (AS1-ATP/AS2-ATP, red line) and active (AS1-ADP/AS2-ATP, cyan line) configurations. The X-axis represents the reaction coordinate RC (bond to be formed): the distance (in Å) between the oxygen atom of the catalytic water and the phosphorous atom of the ATP molecule γ-phosphate group (distance ATP-PG - wat-O). The substrate (S), transition state (TS) and product (P).

ATP hydrolysis facilitates separation of the ATPase heads. Once both active sites are occupied by ADP molecules, the last step in the analysis must necessarily explore the behaviour of the head dimer in this arrangement. From a biochemical point of view, the need for ATP binding and hydrolysis for both DNA loading[27–30] and unloading[31–33] has been described. Also, separation of the ATPase head domain heterodimer, which allows DNA to pass through, is assumed in current models of cohesin function[2,27,29–33,36,51]. However, the underlying details of this mechanism have yet to be reported in order to allow quantification of the contribution of the ATP molecules to the maintenance of the closed conformation of the ring.

To gain insight into this matter, two equivalent structures of the head complex were generated, one containing ATP at both active sites (AS1-ATP/AS2-ATP condition: the system in a conformation prior to the ATP hydrolysis), and the other containing ADP in both active sites (AS1-ADP/AS2-ADP condition: the system after the hydrolysis events). Both structures were stabilized over 150 ns of free MD simulation. For each condition, 5 individual structures were extracted from the free MD trajectories: one every 4 ns from 104 ns to 120 ns. Using those structures, the separation of the heterodimer head domains was analysed via steered MD (SMD) simulations (Fig. 4a and Supplementary Fig. 6), forcing their centres of mass to separate from each other 32.5 Å over 13.0 ns. The values measured for the accumulated work over the 10 SMD trajectories (5 for each condition) were used to estimate the free-energy difference associated with the opening of the head in both conditions, using Jarzynski's equality[52]. Jarzynski's equality allows us to estimate the free-energy difference between two quasi-equilibrium states (in our case, closed and open Smc1A-head/Smc3-head/Rad21-Cter complexes) by collecting the work done over non-equilibrium transitions between those states (in our case, the SMD trajectories).

Force values measured along the head separation in the AS1-ATP/AS2-ATP condition (Fig. 4b, red lines) showed a peak after the first 5 ns (around 3.5 Å of separation between the centres of mass) in all the trajectories.

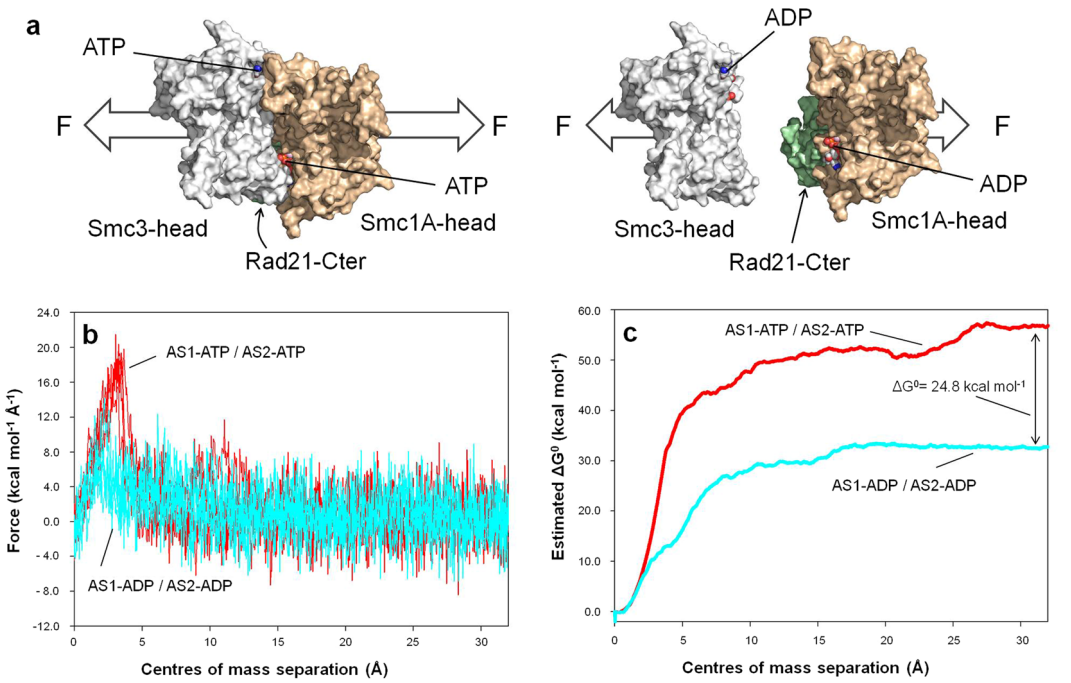


Figure 4. ATP hydrolysis at AS1 and AS2 facilitates head separation. (**a**) Schematic overview of the head separation induced by SMD simulations. The Smc1A-head (brown), Smc3-head (grey) and Rad21-Cter (green) domains are shown and the nucleotide (ATP or ADP) locations in both active sites are indicated. Force (F) direction is marked by white arrows. (**b**) Forces exerted in SMD trajectories over the separation between the centres of mass of Smc1A-head and Smc3-head domains. Points from all trajectories for the AS1-ATP/AS2-ATP condition (red) and for AS1-ADP/AS2-ADP condition (cyan) are shown. (**c**) Estimated free-energy difference (kcal mol^{-1}) over the separation between the centres of mass of the Smc1A-head and Smc3-head domains computed using Jarzynski's equality over 5 SMD trajectories for each condition.

This peak is significantly larger than that observed in the AS1-ADP/AS2-ADP condition (Fig. 4b, cyan lines). In the calculation of the free energy, this force peak represents the largest contribution to the difference between the two conditions: $\Delta G^0 = 24.8$ kcal mol^{-1}. It should be noted that this estimated value is approximately 81% of the average ΔG^0 associated with the hydrolysis of two molecules of ATP in human resting muscle conditions: $\Delta G^0 = 30.6$ kcal mol^{-1} [53], which suggests that this system is highly efficient from an energetic point of view and supports the idea that ATP hydrolysis at both active sites allows head separation.

Pathogenic variants and mutants with an associated phenotypic effect. Our dynamic system at atomic scale of the Smc1A-head and Smc3-head domains provides an advantageous framework for the investigation of pathologies and phenotypic variations associated with specific mutations of residues located in these domains. Table 1 and Figs 5 and 6c summarize this information for some human pathogenic variants as well as for residues whose phenotypic behaviour has been reported in the literature.

The residues affecting ATPase activity at AS1 and AS2 can be grouped into four clusters. The first is composed of the residues Smc1A-N34, Smc1A-R57, Smc3-G1118 and Smc3-Q1119 (depicted in green in Fig. 5a and b). The mutations N34S and R57W in human Smc1A have been related to endometrioid carcinoma[54], with the role of both residues being related to the correct positioning of ATP at AS1. Smc1A-N34 stabilizes the position of the catalytic water during the initial steps of the ATPase reaction (Figs 1c and 2c) and the position of the planar structure of the γ-phosphate during the transition state (Fig. 2c). Smc1A-R57 enters into contact with the α-phosphate group of ATP, thereby stabilizing its position during the entire reaction. Smc3-G1118 and Smc3-Q1119 are located in close contact to Smc1A-R57, allowing for its correct positioning. The mutation of the indicated residues above (depicted in green in Fig. 5a and b) will alter the positioning of ATP at AS1 as well as the progress of the ATPase reaction at this site and subsequent activation of AS2 and head opening.

The second group of residues is composed of the amino acids Smc1A-N1166, Smc3-D1143, Smc3-Q1147 and Smc3-A1148 (depicted in yellow in Fig. 5a and d). The variant residues Smc1A-N1166T and Smc3-Q1147E have been found in patients with CdLS[12,55,56] whereas mutated amino acids Smc3-D1143H and Smc3-A1148T have been related to acute myeloid leukaemia[21,54] and colorectal cancer[54,57], respectively. The mutant Smc3-Q1147E was previously reported to potentially be involved in maintaining the dimerization contact between the Smc1A-head and Smc3-head domains[12]. Interestingly, in the dynamic model, Smc3-Q1147 was found to be involved in correctly locating residues around Smc1A-K1120. To establish whether the mutated Smc3-Q1147E residue can play a differential role in activation of the AS2 site, the same experiment as the previous one illustrated in Fig. 3a was performed but replacing Smc3-Q1147 by Glu. The result (Fig. 5c) indicated that, during the 150 ns trajectory in the presence of the mutant residue, the distance of Smc1A-K1120 from the catalytic water of AS2 was

Protein	Mutation	Disease	Location	References
Smc1A	N34S	Endometroid carcinoma	Active site 1	54
Smc1A	R57W	Endometroid carcinoma	Active site 1	54
Smc1A	V58_R62del	Cornelia de Lange Syndrome	Putative binding to DNA	9, 55, 58, 59
Smc1A	R1090C	Melanoma	Active site 2 (activation)	20, 54
Smc1A	F1122L	Cornelia de Lange Syndrome	Active site 2 (activation)	9, 59, 74
Smc1A	R1123W	Cornelia de Lange Syndrome	Active site 2 (activation)	59, 63
Smc1A	N1166T	Cornelia de Lange Syndrome	Active site 2	55
Smc3	H55Y	Colorectal cancer	Putative binding to DNA	22, 54
Smc3	G1118V	Acute myeloid leukaemia	Active site 1	21, 54
Smc3	Q1119K	Acute myeloid leukaemia	Active site 1	21, 54
Smc3	D1143H	Acute myeloid leukaemia	Active site 2	21, 54
Smc3	Q1147E	Cornelia de Lange Syndrome (CS: moderate*)	Active site 2 (activation)	12, 56
Smc3	A1148T	Colorectal cancer	Active site 2	54, 57

Table 1. Human pathogenic variants. *The patient Smc3-Q1147E showed craniofacial dysmorphism with brachycephaly, arched eyebrows, a depressed nasal bridge and severe ptosis. Additionally, he had heart abnormalities with atrial and septal defects as well as developmental delays and a learning disability[12]. Clinical severity (CS) has been annotated according to Kline *et al.*[58].

intermediate between the non-active and the active structures; it was only compatible with an active arrangement for 0.03% of the total time, in contrast to 13.45% in the case of the wild-type Smc3-Q1147 residue. This therefore predicts greatly reduced (but not completely abrogated) ATPase activity at the AS2 site in the cohesin head of the patient. This finding is very exciting as, to the best of our knowledge, this is the first time that a mutated residue from a CdLS patient has been assigned a specific functional role in a dynamic context involving the cohesin complex.

The third group is composed of the Smc1A amino acids: R1090, F1122 and R1123 (depicted in magenta in Fig. 5a and e). The mutant Smc1A-R1090C has been associated with melanoma[20, 54] and variant residues Smc1A-F1122L[9, 58, 59] and Smc1A-R1123W[59] have been found in CdLS patients. The most interesting fact regarding this group of residues is that their positions in the structure are closely related to the movement of Smc1A-K1120 during the activation of AS2. Drastic mutations such as Arg to Trp, in the case of Smc1A-R1123, or to Cys, in the case of Smc1A-R1090C, or more conservative changes such as Phe to Leu in the case of Smc1A-F1122L, can displace Smc1A-K1120 from its correct positioning at the AS2 site, leading to protein malfunction.

The fourth group of residues is composed of the orthologous positions in human sequences of residues that are mutated in yeast and can bypass the need for Eco1: Smc1A residues L1128, G1131 and D1163 (coloured pink in Fig. 5; equivalent to yeast Smc1 residues L1129, G1132 and D1164, respectively[31]). The position of Smc1A-G1131 and Smc1A-D1163, close to AS2, is compatible with differences in functionality when the Gly residue is mutated to Ser or when the Asp residue is mutated to Glu or Gly[31]. More interesting, however, is the case of Smc1A-L1128, as the mutation of the orthologous Smc1-L1129V in yeast affects the off-rate of cohesins but the equivalent mutation in yeast Smc3-L1126V does not. In our simulations, the Smc1A-L1128 residue stabilizes the hydrocarbon side chain of Smc1A-K1120, thereby allowing its correct location and thus leading the ATPase reaction at AS2. Due to the shorter side chain of Val compared to Leu, such hydrophobic stabilization cannot be maintained in the case of the mutant. In contrast, the equivalent residue Smc3-L1115, although located near AS1, does not play such a central role in the reaction, which explains why the conservative mutation of Leu to Val does not result in a differential phenotype. This asymmetric role of Smc1A-L1128 and Smc3-L1115 in our simulations is in agreement with the differential role of the two equivalent residues in yeast[31].

An additional group of residues of exceptional importance for cohesin function are those related to acetylation-regulated DNA binding. These residues are located both in coiled-coils[37] and in the inner side of the head domains of the Smc1A-Smc3 dimer[2, 32, 33, 38]. During the initial free MD equilibration of the head structure in our model, and possibly due to the lack of constriction forces exerted by the absent coiled coils in the simulation, the relative angle between the head domains grew wider after 40 ns of MD (Fig. 6a), resembling the open structure of the Rad50 head domain associated with DNA[60–62]. To ensure that such movement did not affect the internal structure of either domain, root mean square deviation (rmsd) values were measured during the unrestricted 120 ns MD trajectory of the complex (Fig. 6b). Despite the indicated movement of the head domains, the rmsd values of each domain remained constant (below 3.0 Å). During the relaxation, a number of positively charged residues spontaneously became located in the surface of the dimer (Fig. 6c): Smc1A residues K59, R62 and K149, and Smc3 residues H55, R61, K105, K106 and K157. In addition to the presence of Smc3-K105 and Smc3-K106, already involved in acetylation-regulated DNA binding[2, 32, 33, 38], the presence of three additional amino acids in

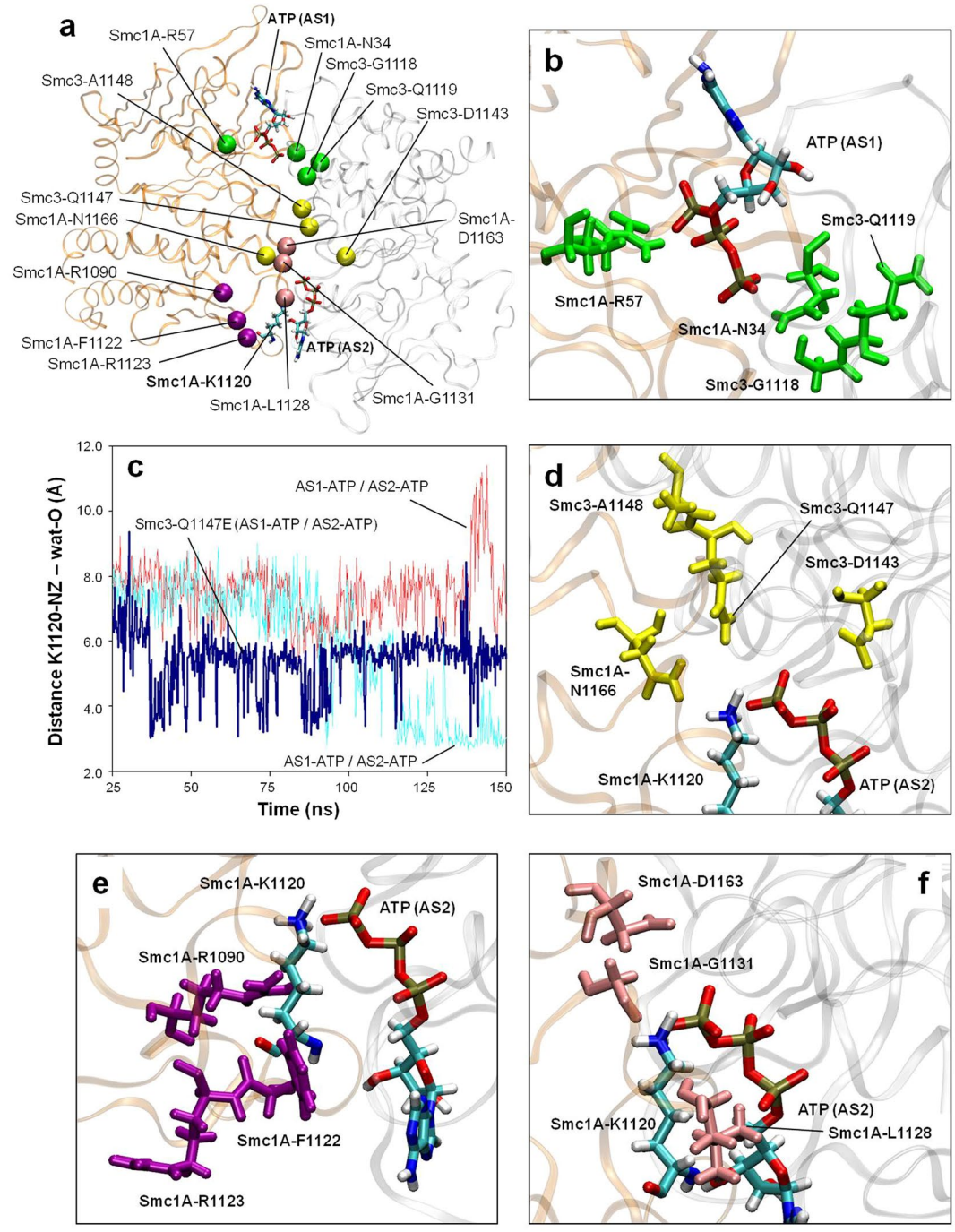


Figure 5. Pathogenic variants. (**a**) Location of the Cα atoms of residues of interest in the neighbourhood of AS1 and AS2. Disease-related variants are shown in green (those affecting AS1), purple (those affecting AS2 activation via Smc1A-K1120 rearrangement) and yellow (those affecting AS2 directly). Residues equivalent to those affected by mutations that bypass the need for Eco1 in yeast are shown in pink. The Smc1A-K1120 residue and both ATP molecules are shown. (**b**) Location of the variants affecting AS1. Residues are depicted in green. (**c**) Evolution of the distance between the oxygen atom of the catalytic water in AS2 and the ε-amino group of the Smc1A-K1120 residue (distance K1120-NZ - wat-O). Distances obtained with wild-type Smc3 prior (red) and after (cyan) ATP hydrolysis at AS1, and distances obtained with the Smc3-N1147E mutant after ATP hydrolysis at AS1 (blue) are shown. (**d**) Location of the variants directly affecting AS2. Residues are depicted in yellow. (**e**) Location of the variants affecting AS2 activation via K1120 rearrangement. Residues are depicted in magenta. The location of K1120 is indicated. (**f**) Location of the residues equivalent to those affected by mutations that bypass the need for Eco1 in yeast. Residues are depicted in pink. The location of K1120 is indicated.

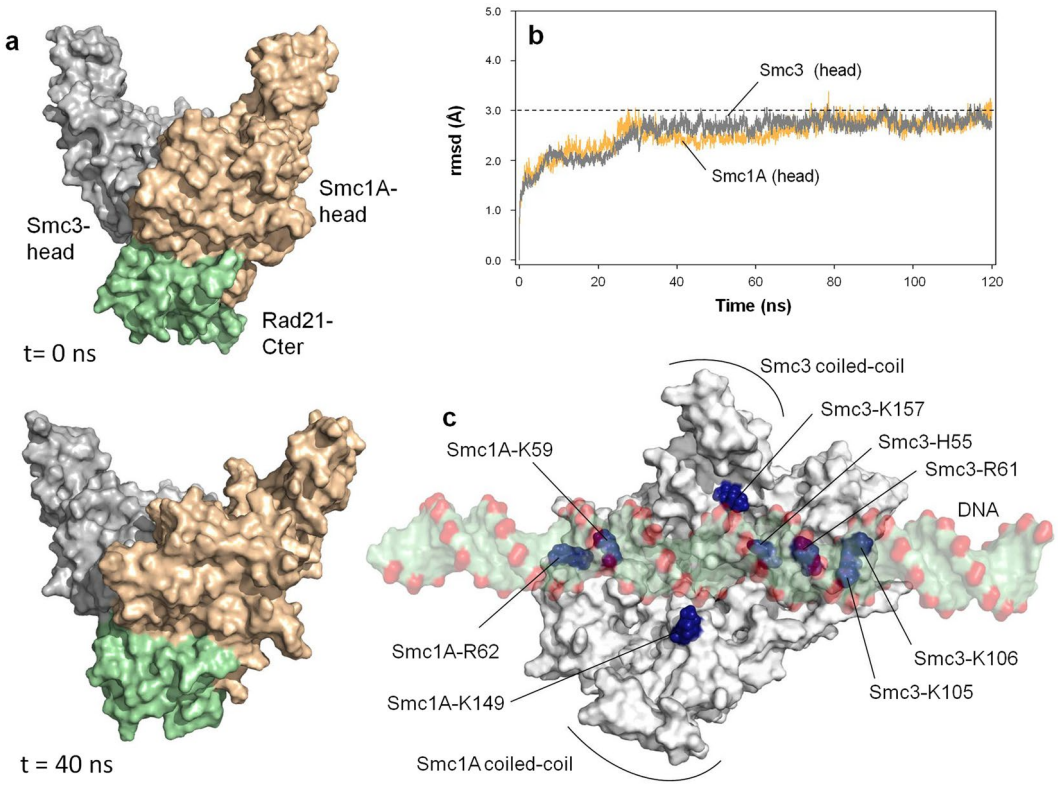


Figure 6. Graphical illustration of a putative interaction of DNA and the head domains of Smc1A and Smc3. (**a**) Relative positions of the Smc1A-head, Smc3-head and Rad21-Cter domains at 0 ns (top) and after 40 ns (bottom) of free MD. (**b**) rmsd values measured over the unrestricted 120 ns MD trajectory of the complex illustrated in **a**. (**c**) Position of positively charged residues in the upper surface of the head complex after 40 ns of MD. The putative position of a DNA molecule, in the equivalent position as the one co-crystallized with Rad50 head domain (PDB code: 5DNY), is indicated.

this group is noteworthy: Smc1A-K59 and Smc1A-R62, deletion of which has been related to CdLS[55, 59, 63] and Smc3-H55, mutation of which to Tyr has been associated with colorectal cancer[22, 54].

Notably, all these positive residues are positioned in a spatial arrangement that is fully compatible with the putative location of a negatively-charged DNA molecule in the surface (Fig. 6c), equivalent to that observed in the case of Rad50[60–62]. This suggests that the initial structure of the dynamic model may mimic the initial position in which not only is the C-ter domain of Rad21 bound to the head domain of Smc1A, but also the positive residues in the head surface of Smc1A and Smc3 are in a position equivalent to the DNA-bound structure; that is, the starting event in the ATPase-dependent opening of cohesin head.

Discussion

Despite the demonstrated relevance of cohesin and cohesin-related proteins to the modulation of important cell functions, the detailed molecular mechanisms underlying the behaviour of the different domains of cohesin proteins has only just begun to be described. In this work, we have analysed the dynamic properties of the human cohesin head domains (Smc1A-head, Smc3-head and Rad21-Cter) at the atomic level using a variety of simulation techniques (free MD, SMD and QM/MM MD). This analysis allows us to infer the role of these domains during ATP hydrolysis events and at the same time to determine how these events affect the atom distribution and function of the protein domains, in a series of events leading to head separation and the subsequent passing of DNA through the open structure (as summarized in Fig. 7).

The first step in the simulation was to describe the role of Rad21 in the hydrolysis of ATP (Fig. 7a). The X-ray structure of the yeast Smc1 head domain bound to the C-terminal domain of Scc1 (the yeast homolog of human Rad21)[5] offered very detailed information relating to this contact, as the surface between the two domains is located in the neighbourhood of AS1. In our system, the simulated protein complex underwent spontaneous rearrangement resulting in the exposure of a group of positive residues in the inner surface of the Smc1A-head/Smc3-head dimer (Fig. 6c). This group includes two Lys residues proposed as involved in acetylation-regulated DNA binding[2, 32, 33, 38], as well as other positive residues mutation of which has been related to CdLS[55, 59, 63] and cancer[22, 54]. In short, the starting point of our simulation corresponds to the head structure bound to Rad21-Cter, in a position compatible with the DNA-bound condition. Our QM/MM MD free-energy analysis for ATP hydrolysis at AS1 provides a quantitative description of the functional role of the Rad21-Cter domain in the cohesin head. The binding of Rad21-Cter allows progression of the ATPase reaction at AS1 by reducing the free-energy barrier by 14.0 kcal mol^{-1}. In our study of this reaction we have also determined the geometry of the residues in

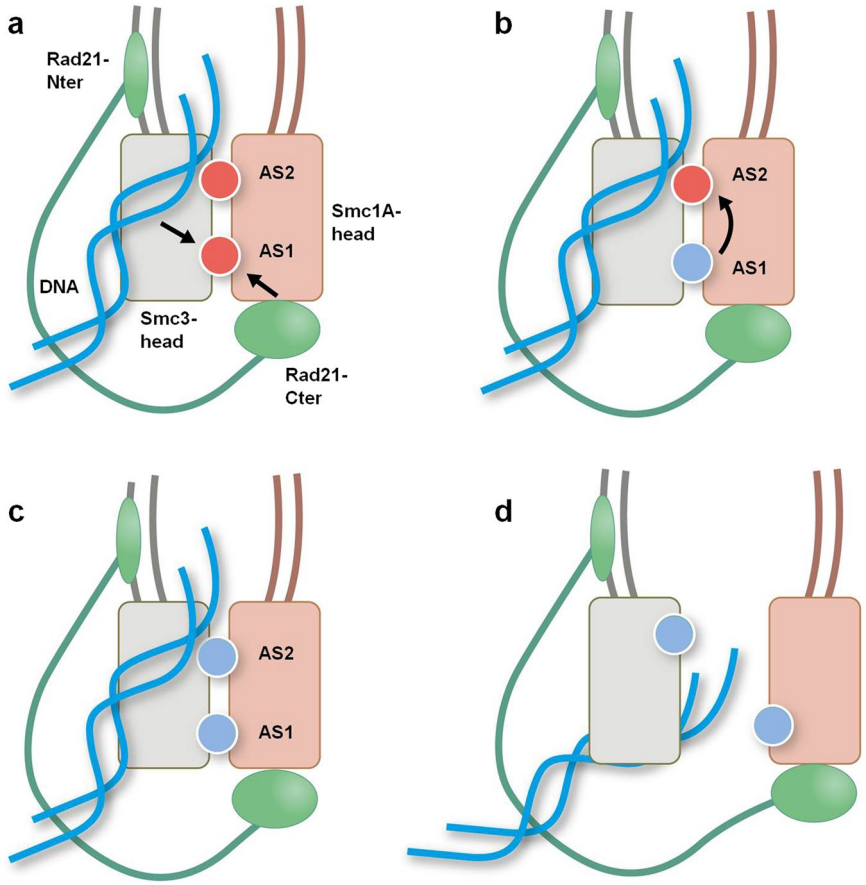


Figure 7. Schematic model for ATP hydrolysis-driven head opening. (**a**) The Rad21-Cter domain binding to the Smc1A-head domain allows hydrolysis at AS1. (**b**) ATP hydrolysis at AS1 induces AS2 activation via Smc1A-K1120 rearrangement. (**c**) ATP hydrolysis takes place at AS2. (**d**) ATP hydrolysis at both active sites facilitates the separation of the ATPase head domains.

the active site during critical events, such as the location of the catalytic water in a first intermediate transition state or stabilization of the ATP gamma phosphate group in the planar transition state (see Supplementary Fig. 1 and Supplementary Video 1 for more details).

Once the ATP hydrolysis reaction has taken place at AS1, an intriguing question is whether hydrolysis at one active site can stimulate the ATPase reaction at the other, a common mechanism of ATP-binding cassette-ATPases (ABC ATPases[64]). Our analysis reveals a sequential structural change, after the hydrolysis event at AS1, that connects the two sites (Fig. 7b), in agreement with the fact that the two ATPase sites are asymmetric[31,51,65]. Although the primary "driving force" leading to the changes in AS2 is still elusive, it is notorious the fact that the variants implicated in CdLS and cancer are roughly located in the pathway that connects AS1 to AS2 (Fig. 5a). We find Smc1A-K1120 to be a major actor in this process. Smc1A-K1120 is conserved in all Smc1A and Smc1B sequences, as well as in the majority of SMC proteins. Interestingly, an exception to this observation is the condensin subunit Smc2, where the Lys residue in this position is replaced by Thr or Ile in different organisms (Supplementary Fig. 7). In condensin, the Smc2 protein plays the equivalent role to Smc3 in cohesin[1,8]. Supposing that the behaviour of AS1 and AS2 is similar in cohesin and condensin dimers, then a Lys residue in this position of Smc2 is not necessary for the ATPase activity. Asymmetrically, the Lys residue in the condensin Smc4 subunit (human Smc4-K1183, equivalent to Smc1A-K1120 in cohesin) is indeed conserved. Our QM/MM MD free-energy analysis of the ATPase reaction at AS2 shows that the new structure of the active centre after the movement of Smc1A-K1120 to a position close to the catalytic water results in a very important reduction of the free-energy barrier. As in the case of AS1, our QM/MM MD investigation also provides a detailed description of the hydrolisis reaction at AS2, including all the intermediate steps (see Supplementary Fig. 5 and Supplementary Video 2), from the initial location of the catalytic water molecule to the stabilization of the transition state and the formation of the final product.

The analysis of protein features linked to the structure of the transition state is a key step in the design of transition state analogues as powerful enzymatic inhibitors[41–43]. In this study, using QM/MM MD techniques, we have obtained a detailed description of AS1 and AS2 during the main transition state as well as during intermediate transition states of the ATPase reactions. The geometry of the active centres of the cohesin head in these transient states is extremely valuable information for future drug development strategies.

After the ATP hydrolysis events at AS1 and AS2, we analysed the subsequent head separation (Fig. 7c), including its quantification in terms of free energy. The free-energy difference for the separation of the Smc1A-head and Smc3-head domains, calculated for the ATP/ATP or ADP/ADP conditions, confirms the key role of the two ATP molecules for the stability of the complex. The data suggest that this process is highly efficient from an energetic point of view and support the hypothesis that the hydrolysis of ATP is followed by the opening of the head. This is in agreement with recent observations of a modelled dimer of the yeast Smc1/Smc3 head[36] as well as recent structural studies of the homologous bacterial SMC dimer[66].

Finally, the dynamic model generated by mixing different (quantum and classical) simulation strategies at atomic scale is a useful framework within which to rationalize the effect of specific mutations involved in both CdLS and cancer; some for the very first time. In particular, the effect of the variant Smc3-Q1147E, found in a CdLS patient[12,55,56] has been analysed in detail, offering a functional explanation for the impaired behaviour of the protein and linking the change at this position to a defect in the activation of AS2 after activation of AS1. The highly reduced, but not completely inhibited, functionality of the complex could justify the clinical findings in this patient who was phenotypically classified as *moderate* according to Kline *et al.*[58].

Our study suggests a functional role in DNA binding of three positively charged residues (Smc1A-K59, Smc1A-R62 and Smc3-H55) mutations of which have been related to CdLS[9,55,58,59] or colorectal cancer[22,54]. In addition, based on our computational approach, we propose a dynamic explanation for several mutants found in yeast that bypass the need for Eco1[31]. Very interestingly, the close interaction of human Smc1A-L1128 with the key residue Smc1A-K1120 and the lack of a symmetrical interaction in the case of the equivalent residue Smc3-L1115 offer a plausible explanation for the asymmetrical effect found[31] between the phenotypes of the orthologous yeast mutants Smc1-L1129V and Smc3-L1126V, respectively.

Altogether, our results reveal the underlying atomic mechanisms of the human Smc1A-head/Smc3-head/Rad21-Cter complex, explaining in detail: (i) the functional role of Rad21-Cter in the activation of AS1; (ii) the modifications that link ATP hydrolysis at AS1 with activation of ATP hydrolysis at AS2; and (iii) the role of ATP hydrolysis in the separation of the heads, in a quantitative manner. This computational approach, mixing quantum and classical simulation techniques, has proved to be a very useful tool for the investigation of phenotypic variants in yeast experiments, to analyse and predict the effect of variants and mutants related to CdLS and cancer, and for use in the future design of therapies and drugs.

Methods

Structure modelling. The three-dimensional model of the complex formed by the human Smc1A-head, Smc3-head and Rad21-Cter domains was built through modelling procedures using the initial protein sequences contained in the UniProtKB database. They are the Smc1A-head: SMC1A_HUMAN (UniProt code: Q14683, residues 1 to 175 and 1058 to 1223); Smc3-head: SMC3_HUMAN (UniProt code: Q9UQE7, residues 1 to 179 and 1045 to 1206); and Rad21-Cter: RAD21_HUMAN (UniProt code: O60216, residues 543 to 629). Three scaffold structures were combined to accurately reproduce various structural features. Rad21-Cter, the active sites (AS1 and AS2) and the interaction interfaces were modelled on the structure of a *Saccharomyces cerevisiae* homodimeric Smc1 ATPase head complex bound to the C-terminal domain of the yeast Scc1(Rad21 orthologue) (Protein Data Bank ID: 1W1W[5]). The structure of Smc3-head was determined by the 3D structure of the *Saccharomyces cerevisiae* Smc3 monomer bound to the Scc1 N-terminal domain (PDB ID: 4UX3[7]). The model of the complex is compatible with recent structures of human cohesin head, obtained by using high-resolution electron microscopy[67], and bacterial SMC head, obtained by crystallography[66]. Multiple sequence alignment of the modelled sequences can be found in the Supplementary Information (Supplementary Fig. 8). The positions of residues around the active sites as well as those of crystallographic water molecules were also refined using the 3D structure of the *Pyrococcus furiosus* Smc homodimer (PDB ID: 1XEX[34]). The ATPγS molecules present at 1 W1W active sites were replaced by either ATP or ADP. The model of human variant Smc3-Q1147E was obtained by replacing the apical amide group in the Smc3-Q1147 residue by a carboxylate group.

Free Molecular Dynamics simulations. Prior to any other simulations, the complex was thermalized and stabilized with free MD simulations using the AMBER14 molecular dynamics package[46]. The 3D structures were solvated with periodic cuboid pre-equilibrated solvent boxes of TIP3P model water molecules[68] using the LEaP module of AMBER, with 12 Å as the shortest distance between any atom in the protein and the periodic box boundaries. Protonation states were determined using the H++ web server (http://biophysics.cs.vt.edu/H++)[69] and Na^+ counterions were added to neutralize the charge of the systems (Smc1A-head/Smc3-head/AS1-ATP/AS2-ATP: 5 Na^+ counterions; Smc1A-head/Smc3-head/Rad21-Cter/AS1-ATP/AS2-ATP: 3 Na^+ counterions; Smc1A-head/Smc3-head/Rad21-Cter/AS1-ADP/AS2-ATP: 2 Na^+ counterions; Smc1A-head/Smc3-head/Rad21-Cter/AS1-ADP/AS2-ADP: 1 Na^+ counterion). All the free MD simulations were performed in the NPT (constant temperature, constant pressure) ensemble, using the PMEMD program of AMBER and the parm99 force field[46]. The SHAKE algorithm was used, allowing a time step of 2 fs.

The different systems used in the simulations were initially relaxed over 15,000 steps of energy minimization with a cut-off of 12 Å. Then the MD simulations were started with a 20 ps heating phase in which the temperature was raised from 0 to 300 K in 10 temperature change steps, after each of which velocities were reassigned. During minimization and heating, the Cα trace dihedrals were restrained with a force constant of 500 kcal mol^{-1} rad^{-2} and gradually released in an equilibration phase in which the force constant was gradually reduced to 0 over 200 ps. After the equilibration phase, 120 to 150 ns of productive MD simulations were obtained for all the systems.

As the Smc1A-head and Smc3-head domains are formed by the N-terminal and C-terminal segments of the two proteins, all the structures showed gaps where the large coiled-coil regions cannot be accurately modelled. Therefore, these gaps were protected by distance restraints to prevent artificial unfolding of these regions.

During long MD trajectories, and in order to improve the sampling of catalytic configurations at both active sites, the position of the catalytic water molecules was maintained in a geometry compatible with hydrolysis, restraining the distance between the oxygen atom of the catalytic water and the phosphorous atom of the ATP γ-phosphate group below 3.5 Å. The angle between these same two atoms and the oxygen atom of the ATP beta phosphate group was kept between 160° and 180°. Both distance and angle restraints were defined using a flat-bottomed potential, allowing free movement within the restrained range. These restraints were released prior to QM/MM MD simulations of the active centres.

QM/MM MD simulations. QM/MM MD simulations were performed using the recently developed Fireball/Amber method[44,45]: a combination of the AMBER molecular dynamics package[46] and Fireball, a local-orbital density-functional theory molecular dynamics technique[47,48]. Two regions (QM and MM) were defined. The MM region was treated in the same manner as in the free MD simulations detailed previously; while the QM region was described using Fireball, with a basis set of optimized numerical atomic-like orbitals (NAOs) with a single *s* orbital for H, *sp3* orbitals for C, N and O, and *sp3d5* orbitals for P, as used in previous works[44,70]. The time step during these calculations was 0.5 fs. The initial structures and initial velocities used in the QM/MM MD simulations were taken from the free MD simulations after they became stable.

Free-energy 2D maps obtained using QM/MM MD simulations were generated as described in the literature[44,70,71]. The conformational space was sampled with long SMD trajectories along the chosen reaction coordinates, generating 7.6×10^6 structures with their associated reaction coordinates and energy values. The QM energy values were distributed in groups of $\sim 1.5 \times 10^4$ different structures on average, across a uniform grid defined by the two reaction coordinates. The partition function was calculated for each group in order to estimate a free-energy surface that was then smoothed via a 3D LOESS local regression. Reaction paths and energy profiles were calculated using MEPSA[49].

1D free-energy profiles were generated by sampling initial structures along the reaction coordinate via SMD. These structures were then relaxed for 5 ps by keeping the reaction coordinate fixed at the corresponding value. Velocities were reassigned every 0.5 ps. The last 0.5 ps of each relaxation was used to estimate the free-energy profile, yielding 7.7×10^4 structures in total, with their associated reaction coordinates and energy values. The QM energy values were distributed in groups of $\sim 10^3$ different structures on average, along uniform 1D grid defined by the reaction coordinate. The partition function was calculated for each group in order to estimate a free-energy profile that was then smoothed via 2D LOESS local regression.

Error analysis was performed for 2D maps and 1D profiles using bootstrap resampling (100 replicates) on the data. In all relevant positions, the standard deviation was found below 0.8 kcal mol^{-1} for 2D maps and below 0.5 kcal mol^{-1} for 1D profiles. Standard deviation representation for 2D free-energy surface of ATP hydrolysis at AS1 in the presence of Rad21-Cter and for 1D free-energy profile of ATP hydrolysis at AS2 in the AS1-ADP/AS2-ATP condition can be found in the Supplementary Information (Supplementary Fig. 9).

Free-energy difference calculations from trajectories of head domain separation. In order to compare the separation of Smc1A-head from Smc3-head, either ATP binding (AS1-ATP/AS2-ATP condition) or ADP binding (AS1-ADP/AS2-ADP condition) free-energy calculations from SMD trajectories were performed using Jarzynski's equality[52]. Five individual SMD trajectories were generated for each condition, taking the ten initial structures from stable free MD trajectories. To prevent protein-protein collisions through the periodic boundaries, these were expanded and the water box was consequently enlarged with pre-equilibrated solvent cuboid boxes of TIP3P water model molecules. To ensure good thermalization of the system after the modification of the water box, a heating protocol similar to that used in the free MD simulations was applied. During each SMD trajectory, the centres of mass of Smc1A-head and Smc3-head were forced to separate 32.5 Å at a constant velocity over 13 ns (2.5 Å ns^{-1}) with a spring constant of 5 kcal mol^{-1} Å^{-2}, in the range of conditions used in similar SMD studies[72,73]. The separation distance was kept constant for 0.1 ns at the start and end of each trajectory to better describe the quasi-equilibrium states at both ends. To avoid large rearrangements of the head structures during SMD, the Cα trace dihedrals were restrained with a 500 kcal mol^{-1} rad^{-2} force constant. The gap protection restraints used in the free MD simulations were also kept in the SMD simulations. For each calculation step, the distance between the centres of mass was recorded to later reconstruct the forces and work generated along each trajectory. The initial distance between the centres of mass was taken as the origin (0.0 Å) of separation in Fig. 4b and c.

References

1. Haering, C. H. & Gruber, S. SnapShot: SMC Protein Complexes Part I. *Cell* **164**, 326–326 e321, doi:10.1016/j.cell.2015.12.026 (2016).
2. Uhlmann, F. SMC complexes: from DNA to chromosomes. *Nat Rev Mol Cell Biol* **17**, 399–412, doi:10.1038/nrm.2016.30 (2016).
3. Rankin, S. & Dawson, D. S. Recent advances in cohesin biology. *F1000Res* **5**, doi:10.12688/f1000research.8881.1 (2016).
4. Gligoris, T. & Lowe, J. Structural Insights into Ring Formation of Cohesin and Related Smc Complexes. *Trends Cell Biol* **26**, 680–693, doi:10.1016/j.tcb.2016.04.002 (2016).
5. Haering, C. H. *et al.* Structure and stability of cohesin's Smc1-kleisin interaction. *Mol Cell* **15**, 951–964, doi:10.1016/j.molcel.2004.08.030 (2004).
6. Huis in 't Veld, P. J. *et al.* Characterization of a DNA exit gate in the human cohesin ring. *Science* **346**, 968–972, doi:10.1126/science.1256904 (2014).
7. Gligoris, T. G. *et al.* Closing the cohesin ring: structure and function of its Smc3-kleisin interface. *Science* **346**, 963–967, doi:10.1126/science.1256917 (2014).
8. Haering, C. H. & Gruber, S. SnapShot: SMC Protein Complexes Part II. *Cell* **164**, 818 e811, doi:10.1016/j.cell.2016.01.052 (2016).
9. Liu, J. & Krantz, I. D. Cornelia de Lange syndrome, cohesin, and beyond. *Clin Genet* **76**, 303–314, doi:10.1111/j.1399-0004.2009.01271.x (2009).
10. Mehta, G. D., Kumar, R., Srivastava, S. & Ghosh, S. K. Cohesin: functions beyond sister chromatid cohesion. *FEBS Lett* **587**, 2299–2312, doi:10.1016/j.febslet.2013.06.035 (2013).

11. Pie, J. *et al.* Special cases in Cornelia de Lange syndrome: The Spanish experience. *Am J Med Genet C Semin Med Genet* **172**, 198–205, doi:10.1002/ajmg.c.31501 (2016).
12. Gil-Rodriguez, M. C. *et al. De novo* heterozygous mutations in SMC3 cause a range of Cornelia de Lange syndrome-overlapping phenotypes. *Hum Mutat* **36**, 454–462, doi:10.1002/humu.22761 (2015).
13. Watrin, E., Kaiser, F. J. & Wendt, K. S. Gene regulation and chromatin organization: relevance of cohesin mutations to human disease. *Curr Opin Genet Dev* **37**, 59–66, doi:10.1016/j.gde.2015.12.004 (2016).
14. Ramos, F. J. *et al.* Clinical utility gene card for: Cornelia de Lange syndrome. *Eur J Hum Genet* **23**, doi:10.1038/ejhg.2014.270 (2015).
15. Chetaille, P. *et al.* Mutations in SGOL1 cause a novel cohesinopathy affecting heart and gut rhythm. *Nat Genet* **46**, 1245–1249, doi:10.1038/ng.3113 (2014).
16. Izumi, K. *et al.* Germline gain-of-function mutations in AFF4 cause a developmental syndrome functionally linking the super elongation complex and cohesin. *Nat Genet* **47**, 338–344, doi:10.1038/ng.3229 (2015).
17. Mannini, L., Menga, S. & Musio, A. The expanding universe of cohesin functions: a new genome stability caretaker involved in human disease and cancer. *Hum Mutat* **31**, 623–630, doi:10.1002/humu.21252 (2010).
18. Pan, X. W. *et al.* SMC1A promotes growth and migration of prostate cancer *in vitro* and *in vivo*. *Int J Oncol* **49**, 1963–1972, doi:10.3892/ijo.2016.3697 (2016).
19. Solomon, D. A., Kim, J. S. & Waldman, T. Cohesin gene mutations in tumorigenesis: from discovery to clinical significance. *BMB Rep* **47**, 299–310, doi:10.5483/BMBRep.2014.47.6.092 (2014).
20. Krauthammer, M. *et al.* Exome sequencing identifies recurrent somatic RAC1 mutations in melanoma. *Nat Genet* **44**, 1006–1014, doi:10.1038/ng.2359 (2012).
21. Metzeler, K. H. *et al.* Spectrum and prognostic relevance of driver gene mutations in acute myeloid leukemia. *Blood* **128**, 686–698, doi:10.1182/blood-2016-01-693879 (2016).
22. Seshagiri, S. *et al.* Recurrent R-spondin fusions in colon cancer. *Nature* **488**, 660–664, doi:10.1038/nature11282 (2012).
23. Hill, V. K., Kim, J. S. & Waldman, T. Cohesin mutations in human cancer. *Biochim Biophys Acta* **1866**, 1–11, doi:10.1016/j.bbcan.2016.05.002 (2016).
24. Williams, M. S. & Somervaille, T. C. Leukemogenic Activity of Cohesin Rings True. *Cell Stem Cell* **17**, 642–644, doi:10.1016/j.stem.2015.11.008 (2015).
25. De Koninck, M. & Losada, A. Cohesin Mutations in Cancer. *Cold Spring Harb Perspect Med* (in press); doi:10.1101/cshperspect.a026476 (2016).
26. Bajic, V., Spremo-Potparevic, B., Zivkovic, L., Isenovic, E. R. & Arendt, T. Cohesion and the aneuploid phenotype in Alzheimer's disease: A tale of genome instability. *Neurosci Biobehav Rev* **55**, 365–374, doi:10.1016/j.neubiorev.2015.05.010 (2015).
27. Ladurner, R. *et al.* Cohesin's ATPase activity couples cohesin loading onto DNA with Smc3 acetylation. *Curr Biol* **24**, 2228–2237, doi:10.1016/j.cub.2014.08.011 (2014).
28. Murayama, Y. & Uhlmann, F. Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature* **505**, 367–371, doi:10.1038/nature12867 (2014).
29. Arumugam, P. *et al.* ATP hydrolysis is required for cohesin's association with chromosomes. *Curr Biol* **13**, 1941–1953, doi:10.1016/j.cub.2003.10.036 (2003).
30. Weitzer, S., Lehane, C. & Uhlmann, F. A model for ATP hydrolysis-dependent binding of cohesin to DNA. *Curr Biol* **13**, 1930–1940, doi:10.1016/j.cub.2003.10.030 (2003).
31. Elbatsh, A. M. *et al.* Cohesin Releases DNA through Asymmetric ATPase-Driven Ring Opening. *Mol Cell* **61**, 575–588, doi:10.1016/j.molcel.2016.01.025 (2016).
32. Beckouet, F. *et al.* Releasing Activity Disengages Cohesin's Smc3/Scc1 Interface in a Process Blocked by Acetylation. *Mol Cell* **61**, 563–574, doi:10.1016/j.molcel.2016.01.026 (2016).
33. Murayama, Y. & Uhlmann, F. DNA Entry into and Exit out of the Cohesin Ring by an Interlocking Gate Mechanism. *Cell* **163**, 1628–1640, doi:10.1016/j.cell.2015.11.030 (2015).
34. Lammens, A., Schele, A. & Hopfner, K. P. Structural biochemistry of ATP-driven dimerization and DNA-stimulated activation of SMC ATPases. *Curr Biol* **14**, 1778–1782, doi:10.1016/j.cub.2004.09.044 (2004).
35. Arumugam, P., Nishino, T., Haering, C. H., Gruber, S. & Nasmyth, K. Cohesin's ATPase activity is stimulated by the C-terminal Winged-Helix domain of its kleisin subunit. *Curr Biol* **16**, 1998–2008, doi:10.1016/j.cub.2006.09.002 (2006).
36. Huber, R. G. *et al.* Impairing Cohesin Smc1/3 Head Engagement Compensates for the Lack of Eco1 Function. *Structure* **24**, 1991–1999, doi:10.1016/j.str.2016.09.001 (2016).
37. Kulemzina, I. *et al.* A Reversible Association between Smc Coiled Coils Is Regulated by Lysine Acetylation and Is Required for Cohesin Association with the DNA. *Mol Cell* **63**, 1044–1054, doi:10.1016/j.molcel.2016.08.008 (2016).
38. Chao, W. C. *et al.* Structural Basis of Eco1-Mediated Cohesin Acetylation. *Sci Rep* **7**, 44313, doi:10.1038/srep44313 (2017).
39. Li, J., Feng, W., Chen, L. & He, J. Downregulation of SMC1A inhibits growth and increases apoptosis and chemosensitivity of colorectal cancer cells. *J Int Med Res* **44**, 67–74, doi:10.1177/0300060515600188 (2016).
40. Wang, J. *et al.* Role of SMC1A overexpression as a predictor of poor prognosis in late stage colorectal cancer. *BMC Cancer* **15**, 90, doi:10.1186/s12885-015-1085-4 (2015).
41. De Vivo, M. Bridging quantum mechanics and structure-based drug design. *Front Biosci (Landmark Ed)* **16**, 1619–1633, doi:10.2741/3809 (2011).
42. Schramm, V. L. Transition States, analogues, and drug development. *ACS Chem Biol* **8**, 71–81, doi:10.1021/cb300631k (2013).
43. Schramm, V. L. Transition States and transition state analogue interactions with enzymes. *Acc Chem Res* **48**, 1032–1039, doi:10.1021/acs.accounts.5b00002 (2015).
44. Mendieta-Moreno, J. I. *et al.* A Practical Quantum Mechanics Molecular Mechanics Method for the Dynamical Study of Reactions in Biomolecules. *Adv Protein Chem Struct Biol* **100**, 67–88, doi:10.1016/bs.apcsb.2015.06.003 (2015).
45. Mendieta-Moreno, J. I. *et al.* Fireball/amber: An Efficient Local-Orbital DFT QM/MM Method for Biomolecular Systems. *J Chem Theory Comput* **10**, 2185–2193, doi:10.1021/ct500033w (2014).
46. Case, D. A. *et al.* AMBER 14. University of California, San Francisco http://ambermd.org/ (2014).
47. Lewis, J. P. *et al.* Further developments in the local-orbital density-functional-theory tight-binding method. *Phys Rev. B* **64**, 195103, doi:10.1103/PhysRevB.64.195103 (2001).
48. Lewis, J. P. *et al.* Advances and applications in the FIREBALL *ab initio* tight-binding moleculardynamics formalism. *Phys. Status Solidi B* **248**, 1989–2007, doi:10.1002/pssb.201147259 (2011).
49. Marcos-Alcalde, I., Setoain, J., Mendieta-Moreno, J. I., Mendieta, J. & Gomez-Puertas, P. MEPSA: minimum energy pathway analysis for energy landscapes. *Bioinformatics* **31**, 3853–3855, doi:10.1093/bioinformatics/btv453 (2015).
50. Hayashi, S. *et al.* Molecular mechanism of ATP hydrolysis in F1-ATPase revealed by molecular simulations and single-molecule observations. *J Am Chem Soc* **134**, 8447–8454, doi:10.1021/ja211027m (2012).
51. Yu, H. Magic Acts with the Cohesin Ring. *Mol Cell* **61**, 489–491, doi:10.1016/j.molcel.2016.02.003 (2016).
52. Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys Rev Lett* **78**, 2690, doi:10.1103/PhysRevLett.78.2690 (1997).
53. Wackerhage, H. *et al.* Recovery of free ADP, Pi, and free energy of ATP hydrolysis in human skeletal muscle. *J Appl Physiol (1985)* **85**, 2140–2145 (1998).
54. Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**, D805–811, doi:10.1093/nar/gku1075 (2015).

55. Gervasini, C. *et al*. Cornelia de Lange individuals with new and recurrent SMC1A mutations enhance delineation of mutation repertoire and phenotypic spectrum. *Am J Med Genet A* **161A**, 2909–2919, doi:10.1002/ajmg.a.36252 (2013).
56. Ansari, M. *et al*. Genetic heterogeneity in Cornelia de Lange syndrome (CdLS) and CdLS-like phenotypes with observed and predicted levels of mosaicism. *J Med Genet* **51**, 659–668, doi:10.1136/jmedgenet-2014-102573 (2014).
57. Mouradov, D. *et al*. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res* **74**, 3238–3247, doi:10.1158/0008-5472.CAN-14-0013 (2014).
58. Kline, A. D. *et al*. Cornelia de Lange syndrome: clinical review, diagnostic and scoring systems, and anticipatory guidance. *Am J Med Genet A* **143A**, 1287–1296, doi:10.1002/ajmg.a.31757 (2007).
59. Mannini, L., Liu, J., Krantz, I. D. & Musio, A. Spectrum and consequences of SMC1A mutations: the unexpected involvement of a core component of cohesin in human disease. *Hum Mutat* **31**, 5–10, doi:10.1002/humu.21129 (2010).
60. Liu, Y. *et al*. ATP-dependent DNA binding, unwinding, and resection by the Mre11/Rad50 complex. *EMBO J* **35**, 743–758, doi:10.15252/embj.201592462 (2016).
61. Schuler, H. & Sjogren, C. DNA binding to SMC ATPases-trapped for release. *EMBO J* **35**, 703–705, doi:10.15252/embj.201694210 (2016).
62. Seifert, F. U., Lammens, K., Stoehr, G., Kessler, B. & Hopfner, K. P. Structural mechanism of ATP-dependent DNA binding and DNA end bridging by eukaryotic Rad50. *EMBO J* **35**, 759–772, doi:10.15252/embj.201592934 (2016).
63. Liu, J. *et al*. SMC1A expression and mechanism of pathogenicity in probands with X-Linked Cornelia de Lange syndrome. *Hum Mutat* **30**, 1535–1542, doi:10.1002/humu.21095 (2009).
64. Hopfner, K. P. Invited review: Architectures and mechanisms of ATP binding cassette proteins. *Biopolymers* **105**, 492–504, doi:10.1002/bip.22843 (2016).
65. Camdere, G., Guacci, V., Stricklin, J. & Koshland, D. The ATPases of cohesin interface with regulators to modulate cohesin-mediated DNA tethering. *Elife* **4**, e11315, doi:10.7554/eLife.11315 (2015).
66. Kamada, K., Su'etsugu, M., Takada, H., Miyata, M. & Hirano, T. Overall Shapes of the SMC-ScpAB Complex Are Determined by Balance between Constraint and Relaxation of Its Structural Parts. *Structure*; doi:10.1016/j.str.2017.02.008 (2017).
67. Hons, M. T. *et al*. Topology and structure of an engineered human cohesin complex bound to Pds5B. *Nat Commun* **7**, 12523, doi:10.1038/ncomms12523 (2016).
68. Jorgensen, W., Chandrasekhar, J., Madura, J., Impey, R. & Klein, M. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935, doi:10.1063/1.445869 (1983).
69. Anandakrishnan, R., Aguilar, B. & Onufriev, A. V. H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* **40**, W537–541, doi:10.1093/nar/gks375 (2012).
70. Mendieta-Moreno, J. I. *et al*. Quantum Mechanics / Molecular Mechanics Free Energy Maps and Nonadiabatic Simulations for a Photochemical Reaction in DNA: Cyclobutane Thymine Dimer. *J Phys Chem Lett* **7**, 4391–4397, doi:10.1021/acs.jpclett.6b02168 (2016).
71. Martin-Garcia, F., Mendieta-Moreno, J. I., Lopez-Vinas, E., Gomez-Puertas, P. & Mendieta, J. The Role of Gln61 in HRas GTP hydrolysis: a quantum mechanics/molecular mechanics study. *Biophys J* **102**, 152–157, doi:10.1016/j.bpj.2011.11.4005 (2012).
72. Cheung, L. S. *et al*. Characterization of monobody scaffold interactions with ligand via force spectroscopy and steered molecular dynamics. *Sci Rep* **5**, 8247, doi:10.1038/srep08247 (2015).
73. Kalyaanamoorthy, S. & Chen, Y. P. A steered molecular dynamics mediated hit discovery for histone deacetylases. *Phys Chem Chem Phys* **16**, 3777–3791, doi:10.1039/c3cp53511h (2014).
74. Kline, A. D. *et al*. Natural history of aging in Cornelia de Lange syndrome. *Am J Med Genet C Semin Med Genet* **145C**, 248–260, doi:10.1002/ajmg.c.30137 (2007).

Acknowledgements

This work is supported by the Spanish MINECO (contracts IPT2011-0964-900000 and SAF2011-13156-E to P.G-P and projects MAT2014-59966-R and "María de Maeztu" Programme for Units of Excellence in R&D, MDM-2014-0377, to J.O.), the Spanish Ministry of Health -ISCIII-*Fondo de Investigación Sanitaria* (FIS) (Ref.# PI15/00707, to F.J.R. and J.P.) and the *Diputación General de Aragón* (Grupo Consolidado B20, European Social Fund "Construyendo Europa desde Aragón", to J.P.). The computational support of the "Centro de Computación Científica - CCC-UAM" is acknowledged. We are grateful to Lucía García-Ledo for her useful comments. We thank José L. Belio for the artwork and Toffa Evans for valuable assistance with the manuscript.

Author Contributions

I.M.-A., J.I.M.-M. and D.S. performed the computational analysis. J.M. designed and supervised the computational analysis. M.C.G.-R., B.P. and M.H.-M. analysed the human variant data. F.J.R. and J.P. supervised and discussed the variant analysis. J.O. discussed the computational results. I.M.-A. and P.G.-P. wrote the manuscript. P.G.-P. supervised the research.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-03118-9

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

C | Video: ATP hydrolysis at AS1 in the presence of Rad21

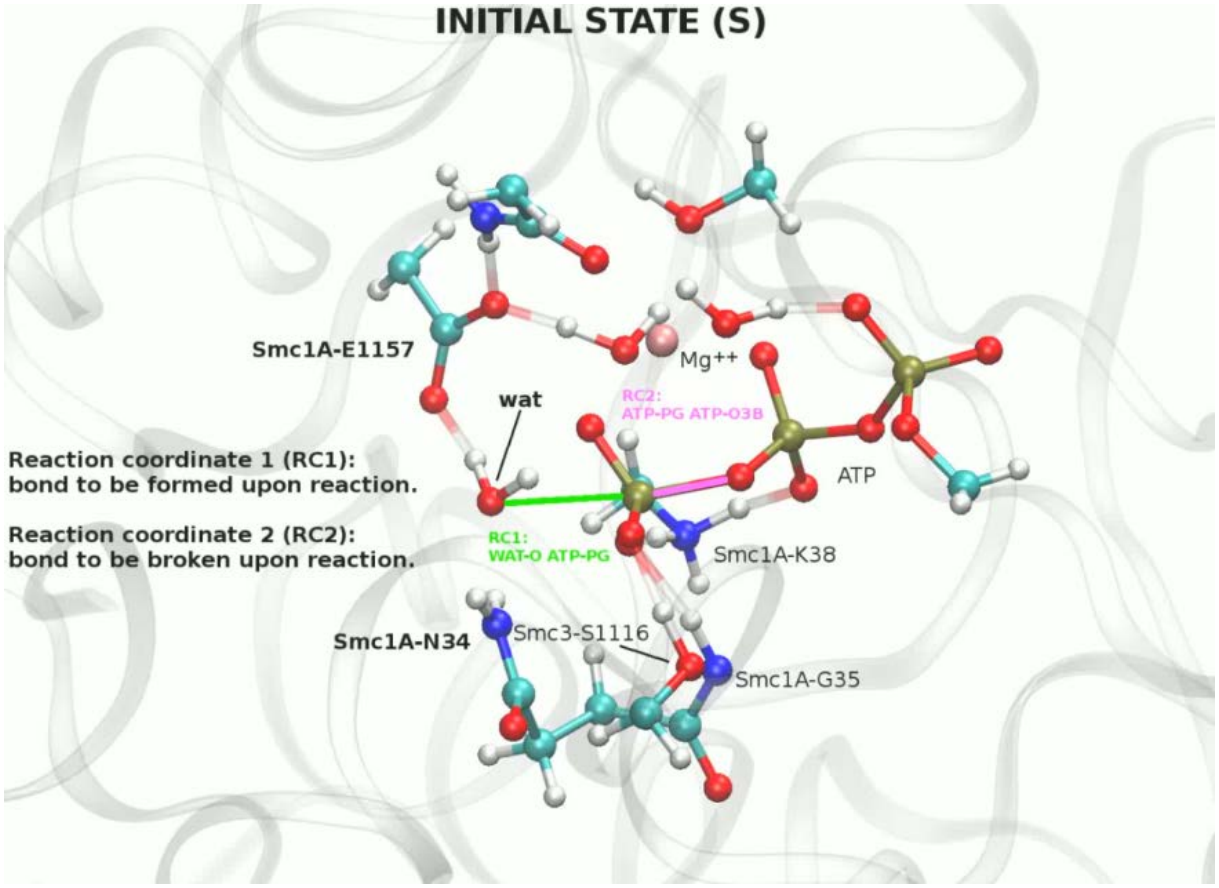


Figure 69: *Video: ATP hydrolysis at AS1 in the presence of Rad21. The movie shows the reaction along the MEPSA minimum energy path that is shown in figure 51. Four steps in the reaction are highlighted: initial structure (S), stabilization of the catalytic water molecule, transition state (TS) and final product (P). Source: Marcos-Alcalde et al. (2017)*[24]*. A copy of the video is present in the DVD that accompanies this thesis and is available on-line at: https://youtu.be/ZsPWPVCUiG8*

D | Video: Positioning of Smc1A-K1120

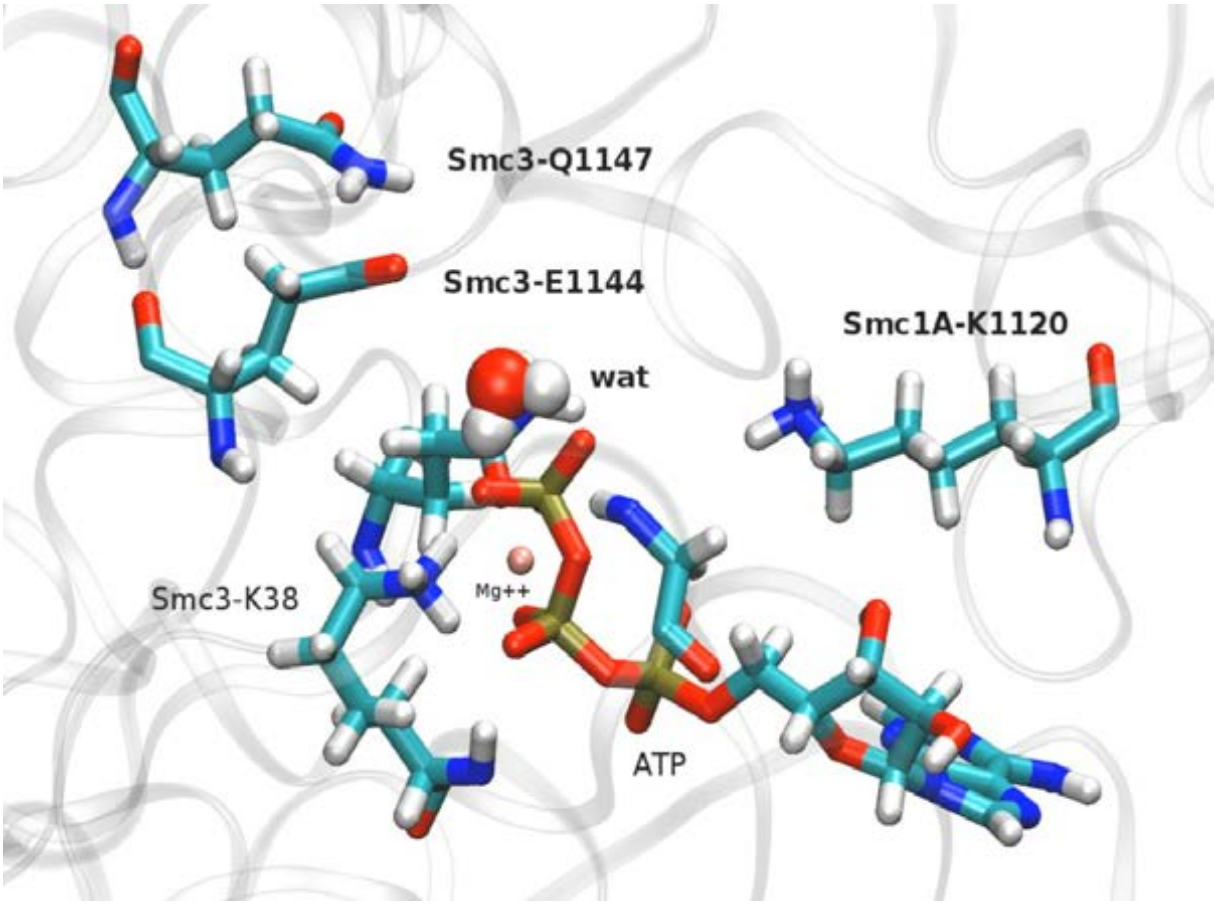


Figure 70: *Video: Positioning of Smc1A-K1120. After ATP hydrolysis at AS1, Smc1A-K1120 moves close to the AS2 catalytic water molecule and remained in its new location in a stable conformation. The movie shows the MD from time 75 to 150 ns of the AS1-ADP/AS2-ATP trajectory discussed in figure 54. Position of residue Smc3-Q1147 is also indicated. Protons are not shown during movement to avoid smoothing artifacts. A copy of the video is present in the DVD that accompanies this thesis and is available on-line at: https://youtu.be/edfNbtQK8ZA*

E | Video: ATP hydrolysis at AS2 in its active form (AS1-ADP/AS2-ATP)

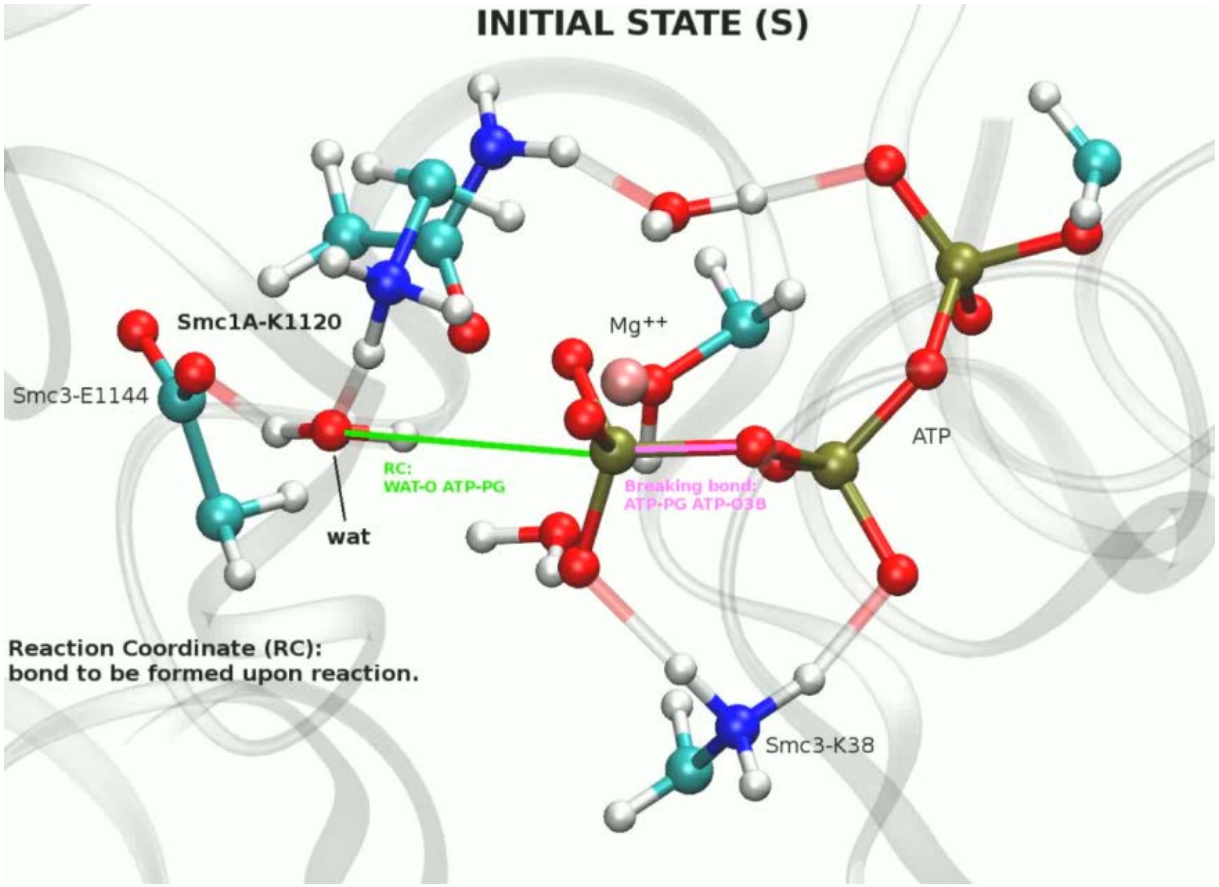


Figure 71: *Video: ATP hydrolysis at AS2 in its active form (AS1-ADP/AS2-ATP). The movie shows the reaction along the RC1 coordinate as indicated in figure 55, corresponding to the free energy profile shown in figure 58. Three steps in the reaction are highlighted: initial structure (S), transition state (TS) and final product (P). A copy of the video is present in the DVD that accompanies this thesis and is available on-line at: https://youtu.be/hVyNMAhOxMo*

F | Python script for Jarzynski calculations

```
#!/usr/bin/env python3.4
import decimal as dec
import numpy as np
import time

###OPTIONS###
input_files_list = [""]
output_files_name = ""
KT = 0.593
precision = 20
#############

##FUNCTIONS##
def calculate_forces (current, target, force_constant):
    forces = np.empty(np.shape(current)[0]);
    for counter in range (np.shape(target)[0]):
        forces[counter] = 2 * (current[counter] - target[counter]) *
                                            force_constant[counter] *
                                            natom

    return forces

def calculate_distance_increments (target):
    distance_increments = np.zeros(np.shape(target)[0]);
    for counter in range (1,np.shape(target)[0]):
        distance_increments[counter] = (target[counter] - target[counter
                                            -1])
    return distance_increments

def calculate_accumulated_work (forces, distance_increments):
    accumulated_work = np.zeros(np.shape(forces)[0]);
    accumulated_work2 = np.zeros(np.shape(forces)[0]);
    for counter in range (1,np.shape(forces)[0]):
        accumulated_work[counter] = accumulated_work[counter-1] + ((
                                            forces[counter-1] + forces[
                                            counter]) *
                                            distance_increments[counter]
                                             * 0.5)
    return accumulated_work

def get_forces_and_accumulated_work (input_file):
    print ("Reading file")
```

```
    target, current, force_constant = np.loadtxt(input_file, unpack=True
                                                 )
    print ("Calculating forces")
    forces = calculate_forces (target, current, force_constant)
    print ("Calculating distance increments")
    distance_increments = calculate_distance_increments (target)
    print ("Calculating accumulated work")
    accumulated_work = calculate_accumulated_work (forces,
                                                   distance_increments)
    return forces,accumulated_work

#--------

def calculate_Jarzynski (accumulated_work_stack, step_counter, B):
    sum_of_exps = 0
    for stack_counter in range (len(accumulated_work_stack)):
        sum_of_exps = dec.Decimal(sum_of_exps + dec.Decimal(-B * dec.
                                                 Decimal(
                                                 accumulated_work_stack[
                                                 stack_counter][step_counter]
                                                 )).exp())
    free_energy_difference = dec.Decimal(dec.Decimal(sum_of_exps / len(
                                                   accumulated_work_stack)).ln() *
                                                   dec.Decimal(-KT))
    return free_energy_difference

def get_Jarzynski (accumulated_work_stack, B):
    print ("Calculating Jarzynski")
    local_start_time = time.time()
    free_energy_difference = np.empty(np.shape(forces_stack[0])[0])
    print_counter = 0
    for step_counter in range (np.shape(forces_stack[0])[0]):
        free_energy_difference[step_counter] = calculate_Jarzynski (
                                                 accumulated_work_stack,
                                                 step_counter, B)
        if (print_counter == 100000):
            remaining_time = (np.shape(forces_stack[0])[0]-step_counter+
                                                     1)*((time.time()-
                                                     local_start_time)/(
                                                     step_counter+1))
            print (('{0:.1f} seconds remaining.').format(remaining_time)
                                                     )
            print_counter = 0
        print_counter += 1
    return free_energy_difference

#--------

def write_output_files (forces_stack, accumulated_work_stack,
                                    free_energy_difference,
                                    output_files_name):
    print ("Writing output files")
    local_start_time = time.time()
    name = output_files_name + "_forces.dat"
    forces_out = open(name, 'w')
    name = output_files_name + "_acc_works.dat"
    accumulated_works_out = open(name, 'w')
    name = output_files_name + "_free_energ_diff.dat"
```

```
    free_energy_difference_out = open(name, 'w')

    forces_string = ""
    accumulated_works_string = ""
    free_energy_difference_string = ""
    print_counter = 0
    for step_counter in range (np.shape(forces_stack[0])[0]):
        for stack_counter in range (len(accumulated_work_stack)):
            forces_string = forces_string + str(forces_stack[
                                                stack_counter][
                                                step_counter]) + " "
            accumulated_works_string =  accumulated_works_string + str(
                                                accumulated_work_stack[
                                                stack_counter][
                                                step_counter]) + " "
        if (print_counter == 100000):
            remaining_time = (np.shape(forces_stack[0])[0]-step_counter+
                                                1)*((time.time()-
                                                local_start_time)/(
                                                step_counter+1))
            print ('{0:.1f} seconds remaining.'.format(remaining_time))
            print_counter = 0
        print_counter += 1
        forces_string = forces_string + "\n"
        accumulated_works_string =  accumulated_works_string + "\n"
        free_energy_difference_string = str(free_energy_difference[
                                            step_counter]) + "\n"
        forces_out.write(forces_string)
        accumulated_works_out.write(accumulated_works_string)
        free_energy_difference_out.write(free_energy_difference_string)
        forces_string = ""
        accumulated_works_string =  ""

    forces_out.close()
    accumulated_works_out.close()
    free_energy_difference_out.close()

####MAIN####

start_time = time.time()

dec.getcontext().prec = precission

B = dec.Decimal(1/dec.Decimal(KT))
number_of_lines = 0

for file_name in input_files_list:
    print ("File " + file_name)
    forces,accumulated_work = get_forces_and_accumulated_work (file_name
                                        )
    try:
        forces_stack.append(forces)
        accumulated_work_stack.append(accumulated_work)
    except:
        forces_stack = [forces,]
        accumulated_work_stack = [accumulated_work,]

free_energy_difference = get_Jarzynski (accumulated_work_stack, B)
```

```
write_output_files (forces_stack, accumulated_work_stack,
                                    free_energy_difference,
                                    output_files_name)
print("--- Total execution time = %s seconds ---" % (time.time() -
                                    start_time))
```